

A UNIFIED THEORY OF CAPITALIST DEVELOPMENT

**Adolfo
Figueroa**



CENTRUM
CENTRO DE NEGOCIOS



About the Author

Adolfo Figueroa received his Doctorate Degree (Ph.D.) in Economics from Vanderbilt University, Nashville, USA, in 1972. He studied under the prominent scientific influence and guidance of Professor Nicholas Georgescu-Roegen, the famous bio-economist.

He has been a faculty member at the Pontificia Universidad Católica del Perú (PUCP) since 1970, first at the Economics Department and then at CENTRUM Católica, Business Graduate School of the PUCP. He is Professor Emeritus of Economics of this University.

Dr. Figueroa's fields of teaching and research include economic development, economic and social inequalities, labor markets, and agricultural development.

Outside his native Peru, he has been visiting professor at several universities in the United States (Illinois at Urbana-Champaign, Notre Dame, Texas at Austin, and Wisconsin at Madison), United Kingdom (University of Oxford), and Latin America (graduate schools in Brazil, Nicaragua, and Ecuador). He has also served as consultant to multilateral and international organizations such as The World Bank, ILO, FAO, IFAD, and IDB.

Dr. Figueroa has widely published his research work in books, papers in academic journals, and in collective volume books. He is the author of *Capitalist Development and the Peasant Economy in Peru*, published by the prestigious Cambridge University Press (1984 and reprinted in paper back in 2008).

**A UNIFIED THEORY OF
CAPITALIST DEVELOPMENT:
GROWTH, INEQUALITY, AND THE ENVIRONMENT**

Adolfo Figueroa, Ph.D.

Professor Emeritus of Economics

Centrum Business School, Catholic University of Peru

Email: afiguer@pucp.edu.pe

August 2012

Author's Note. This is a revised version of my book *A Unified Theory of Capitalist Development* (Buenos Aires: Cengage Learning, 2009), ISBN 978-987-1486-20-5. The conclusions of the book remain, but the scientific rigor has been greatly improved.

Suggested citation:

Figueroa, Adolfo (2009), *A Unified Theory of Capitalist Development* (Buenos Aires: Cengage Learning, 2009), revised an enlarged version, August 2012, available on line at <http://www.freescience.info>

Praise to Adolfo Figueroa, *A Unified Theory of Capitalist Development* (Buenos Aires: Cengage Learning, 2009):

Adolfo Figueroa, a Peruvian economist ...has written a breathtakingly ambitious book. Physics is still struggling to integrate in a single unified theory the insights of general relativity and quantum theory. Figueroa aims not only to build partial models capable of explaining different parts of the capitalist system, but also a unified model from which each of these special cases can be derived...[T]he reader would have to be very blinkered not to agree that Figueroa is on to something important. Figueroa deserves to be read, and in an ideal world this book would become required reading for economics students throughout the world.

Victor Bulmer-Thomas

Chatham House, The Royal Institute of International Affairs, UK

Book Review Article in *Journal of Human Development and Capabilities*, Vol. 11, No.2, May 2010; pp. 359-361.

CONTENTS

List of Tables	8
List of Figures	8
On Notations	9
Acknowledgments	11

INTRODUCTION 12

1. THE RULES OF SCIENTIFIC KNOWLEDGE 24

1.1 Scientific Rules from Popperian Epistemology	24
1.2 The Economic Process	26
1.3 The Alpha-Beta Method	27
1.4 The Alpha-Beta Method in Economics	32

2. PRODUCTION AND DISTRIBUTION UNDER CAPITALISM: EMPIRICAL REGULARITIES 41

2.1 Scope of the Book: Capitalist Countries, 1950-2010	41
2.2 A Brief History of Capitalism and Colonialism	42
2.3 Empirical Regularities on Production and Distribution	44
2.4 The Need of theory to Explain the Empirical Regularities	48

3. STANDARD AND CLASSICAL ECONOMICS 52

3.1 The Neoclassical Society	52
3.2 The Keynesian Society	62
3.3 The Classical Society	68
3.4 Empirical Consistency: The First World Countries	75
3.5 Generalizing the Models to Theories	76

• PART I Partial Theories 81

4. THE FIRST WORLD 82

4.1 Epsilon: Socially Homogeneous Class Society and Under-populated	82
4.2 The Nature of the Labor Market	84
4.3 A Static Model of the Epsilon Theory: The Short Run	88
4.4 Market General Equilibrium	92
4.5 Empirical Predictions	97
4.6 Empirical Consistency: The First World Countries	101

5. THE THIRD WORLD 105

5.1 Omega: Socially Homogeneous Class Society and Overpopulated	105
5.2 A Static Model of the Omega Theory: The Short Run	106
5.3 Market General Equilibrium	111

5.4 Empirical Predictions	116
5.5 Empirical Consistency: Third World Countries with Weak Colonial Legacy	119

6. THE THIRD WORLD WITH COLONIAL LEGACY 123

6.1 Sigma: Socially Hierarchical and Overpopulated Society	123
6.2 A Static Model of the Sigma Theory: The Short Run	124
6.3 General Equilibrium	126
6.4 Empirical Predictions	129
6.5 Empirical Consistency	131

• PART II Towards a Unified Theory of the Capitalist System 136

7. INEQUALITY AND SOCIAL DISORDER 137

7.1 Limited Tolerance to Inequality	137
7.2 Government Behavior towards Inequality	140
7.3 Social Order as Production Factor	144
7.4 General Equilibrium with Social Disorder	146
7.5 Empirical Predictions	148
7.6 Empirical Consistency: The Capitalist World	152
7.7 Conclusions	154

8. INVESTMENT IN PHYSICAL CAPITAL 159

8.1 Aversion to Risky Games Behavior vs. Risk Aversion Behavior	160
8.2 Aggregate Investment Behavior	163
8.3 Empirical Consistency: The Capitalist World	164
8.4 Inequality and International Competition	166
8.5 The Role of Credit Markets	167
8.6 The Role of Insurance Markets	170
8.7 Basic Markets	175

9. EDUCATION AND INVESTMENT IN HUMAN CAPITAL 178

9.1 Static Models with Exogenous Human Capital	178
9.2 Process Analysis of Education	179
9.3 The Role of the Initial Inequality	181
9.4 Transforming Education into Human Capital	183
9.5 Exclusion vs. Wage Discrimination in Labor Markets	185
9.6 Empirical Predictions	186
9.7 Empirical Consistency: The Capitalist World	187
9.8 Conclusions	188

• PART III A Unified Theory of the Capitalist System 192

10. PRODUCTION AND DISTRIBUTION IN THE CAPITALIST SYSTEM 193

10.1 Foundations of the Unified Theory	193
10.2 A Static Model: Explaining Income Level Differences	195
10.3 Within-country Inequalities	198
10.4 The Question of Short Run Changes	198

10.5 The Question of International Trade	200
10.6 Additional Empirical Evidence	201
11. GROWTH AND INEQUALITY IN THE CAPITALIST SYSTEM	205
11.1 A Dynamic Model of the Unified Theory	205
11.2 A Dynamic Epsilon Model	207
11.3 A Dynamic Omega Model	210
11.4 A Dynamic Sigma Model	213
11.5 Dynamic Equilibrium of Growth and Distribution	216
11.6 Empirical Evidence	221
11.7 Conclusions	224
12. GROWTH AND INEQUALITY UNDER ENVIRONMENTAL DISTRESS	232
12.1 Towards an Evolutionary Model of the Unified Theory	232
12.2 Model A: Economic Process with Non-renewable Natural Resources	233
12.3 The Intergenerational Consumption Frontier	235
12.4 Model B: Introducing the Laws of Thermodynamics	238
12.5 Model C: Introducing Substitutability between Funds and Flows	243
12.6 Changes in the Intergenerational Consumption Frontier	244
12.7 Economic Growth as Evolutionary Process	246
12.8 Growth and Quality of Life	250
12.9 Empirical Evidence	251
12.10 Conclusions	253
13. A UNIFIED THEORY OF THE CAPITALIST SYSTEM	260
13.1 A New Economic Theory of Capitalism	260
13.2 The Structure of the Unified Theory	261
13.3 Differences between Unified Theory and Standard Economics	262
13.4 Explaining Overall Inequality in the Capitalist System	264
13.5 Comparisons with Other Theories	266
13.6 The Capitalist System as Sigma Society	270
14. SCIENCE-BASED PUBLIC POLICIES	274
14.1 Growth versus Quality of Life: Today's Big Trade Off	274
14.2 Market or Democracy Failures?	275
14.3 Current Public Policies	279
14.4 Alternative Public Policies	280
14.5 Breaking with History	285
14.6 Conclusions	289
Appendix A	290
Appendix B	293
BIBLIOGRAPHY	296

List of Tables

Table 2.1 Evolution of Countries to Capitalism, 1500-2010	50
Table 2.2 Income Level and Income Inequality in Capitalist Countries	51
Table 6.1 Social Structure of Sigma Society: Race, Class, and Citizenship	134

List of Figures

Figure 1.1 Diagrammatic Representation of a Process Analysis	38
Figure 1.2 The Alpha-Beta Method	39
Figure 1.3 The Alpha-Beta Method in Economics	40
Figure 3.1 Neoclassical General Equilibrium	78
Figure 3.2 Keynesian General Equilibrium: Labor and Money Markets	79
Figure 3.3 Classical General Equilibrium	80
Figure 4.1 Work Effort, Efficiency Unemployment, and Labor Demand Curve	103
Figure 4.2 General Equilibrium in the Open Epsilon Society	104
Figure 5.1 General Equilibrium in the Omega Society	122
Figure 6.1 General Equilibrium in the Sigma Society	135
Figure 7.1 Region of Social Tolerance to Inequality	156
Figure 7.2 General Equilibrium with Social Disorder in the Capitalist Sector	157
Figure 7.3 Relationship between Inequality (D) and Social Disorder (SD) by Types of Capitalist Societies	158
Figure 8.1 Credit Market Equilibrium with Rationing	177
Figure 9.1 Theoretical Relations between Education, Human Capital by social Groups	190
Figure 9.2 Theoretical Relations between Income and Human Capital, by Social Groups	191
Figure 10.1 Income Levels by Types of Capitalism	204
Figure 11.1 Steady State Equilibrium in Epsilon Society	226
Figure 11.2 Output per Worker Dynamic Equilibrium in Epsilon Society	227
Figure 11.3 Output per Worker Dynamic Equilibrium and Transitional Dynamics in Epsilon Society	228
Figure 11.4 Growth in Omega Society	229
Figure 11.5 Growth in Sigma Society	230
Figure 11.6 Growth and Distribution in the Capitalist System	231
Figure 12.1 The intergenerational consumption frontier.....	255
Figure 12.2 Depletion and pollution in the economic process	256
Figure 12.3 Limits to Growth under Evolutionary Dynamics.....	257
Figure 12.4 CO2 concentration levels (part per million) from 1000AD to 1995AD	258
Figure 12.5 Growth and Quality of Life	259

On Notations

General Symbols

α	Set of theoretical propositions, primary assumptions of a theory
β	Set of refutable empirical propositions, logically derived from α
δ	Degree of inequality in individual asset endowments
ε	Epsilon society: Abstract society, socially homogeneous and underpopulated
σ	Sigma society: Abstract society, socially heterogeneous and overpopulated
ω	Omega society: Abstract society, socially homogeneous and overpopulated
Π	Pollution
τ	Rate of growth of technological progress (Chapter 12)
μ	Quantity of natural resources per unit of net output
A	Level of technology in the economy
B	Commodity produced domestically
b	Set of statistical relations derived from empirical data
C	Commodity imported as input of commodity B
c	Technological coefficient of input C per unit of commodity B
D	Degree of income inequality
D_j	Quantity of good j demanded
D_h	Quantity of labor demanded
D_h^*	Maximum quantity of wage employment
D_r^*	Effective quantity of credit demanded
E	Foreign exchange
e	Real investment expenditure as proportion of national income
g	Rate of growth of technology (Chapter 11)
H	Labor services
I	Real net investment
K_b	Stock of physical capital of commodity B
K_b^*	Threshold of physical capital needed to operate as a capitalist firm
K_h	Stock of human capital
K_i	Stock of physical capital as infrastructure
k	Capital labor ratio
L	Quantity of workers employed, as wage earners and self-employed
M	Domestic money
N	Quantity of mineral resources utilized as production input
n	Rate of population growth
O	Degree of social order
P	Total real profits
P'	Total nominal profits
p	Premium rate in the wage rate determination
P_b	Nominal price of good B
P_b^*	International price of good B
P_c	Nominal price of good C
P_c^*	International price of good C
P_e	Nominal exchange rate
P_h	Nominal wage rate
P_h^*	Initial nominal wage rate
p	Premium rate in the market wage rate determination

Q_b	Quantity of net output of good B produced in the capitalist sector
q^*	Threshold of quantity supplied in the credit and insurance markets
R	Bank credit
r	Domestic nominal rate of interest
r^*	International nominal rate of interest
S	Standard deviation
S_h	Quantity of workers supplied
S_j	Quantity of good or service j supplied
S_m	Quantity of money supplied
s	Insurance premium rate in the insurance market
T	Time as historical time
t	Time as mechanical time
U	Total unemployment
u	Unemployment rate
u^*	Threshold of efficiency unemployment rate
V	Total real output in the subsistence sector
v	Average labor productivity in the subsistence sector
v'	Marginal labor productivity in the subsistence sector
W	Total real wage bill
w	Real wage rate
w^*	Threshold of efficiency real wage rate
X_b	Quantity of good B exported
Y	National income or total net output
y	Income per capita, or output per worker
Z	Stock of public goods as a composite good
z	Quantity of mineral resources per unit of gross output
z^*	Terms of international trade

Note on Equations

Equations that are presented as a system have a unique number. In the case of structural equations, endogenous and exogenous variables are separated by a semicolon in each equation. Static equilibrium values of the endogenous variables are marked with zero superscript and dynamic equilibrium with asterisk superscript.

Note on Figures

Curves show the exogenous variables of the model within a bracket and separated by a semicolon. Static equilibrium values of the endogenous variables are marked with zero superscript and dynamic equilibrium with asterisk superscript.

ACKNOWLEDMENTS

In the preparation of this revised version I have received comments and suggestions from colleagues and students, too many to name and thank them properly. The research assistance of Erick Vila, Francisco Pardo, and Javier Vásquez at Centrum Business School, Catholic University of Peru in Lima, is greatly appreciated.

INTRODUCTION

Two centuries of economic growth have shown that the capitalist system is technologically very progressive; socially, however, it is not so. Inequality is one of the persistent features of world capitalism. Why does the capitalist system operate in this way?

The current paradigm in economics has made significant progress in explaining capitalism; however, there are some facts that the paradigm cannot explain, as it will be shown in this book. The current paradigm can be defined as the principles that are contained in the university textbooks that are used in economics courses around the world; consequently, its influence on public policies is enormous. Those principles are based on neoclassical and Keynesian economic theories, one for the long run and the other for the short run analysis of capitalism. Both theories will be called standard economics in this book.

This book presents a new economic theory of the capitalist system. This new theory will be able to explain the facts that standard economic can also explain; but it will also explain those facts that standard economics cannot.

In order to introduce the reader to this new view gradually, the foundations of the new theory will be presented step by step in this introduction. This introduction may then be seen as the *synopsis* of a play about the nature of social relations in a capitalist society, which will then be developed fully in the book.

Chapter 1: The Rules of Scientific Knowledge

Economics is a science, a social science. Economics seeks to explain the economic process, which comprises both production and distribution; that is, production of goods and its distribution among social groups in human societies. Capitalist societies will be the sole object of study in this book. This is the scope of the book.

Economics also needs to define the rules of scientific knowledge. Epistemology is the discipline that studies the logic of scientific knowledge, which then gives justification to the scientific rules utilized. Because there are several epistemologies, knowledge is epistemology dependent. In scientific work, therefore, the epistemology to be used must be made explicit. The Popperian epistemology will be used in this book.

According to Popperian epistemology, three things are needed to explain reality:

- (a) A set of facts about the reality under study, which should take the form of empirical regularities.
- (b) A scientific theory. Scientific explanation requires theory. The goal of theory is to transform reality into an abstract world, selecting only those elements that seem to be essential to the understanding of reality, and thus ignoring

the rest. This selection is made by a set of assumptions. Whether the selection was good or bad is determined by the confrontation of the theory with the reality.

- (c) The theory has been empirically corroborated. Does the abstract world constructed by the theory resemble well the real world? If it does, then the theory is a good approximation and explains reality; if it does not, then the theory fails. The resemblance is solved by confronting the empirical predictions of the theory, derived logically from its assumptions, against the set of empirical regularities. If there are no empirical regularities or no theory, then scientific knowledge cannot be achieved.

This chapter shows that these principles are also applicable to the study of the economic process. A particular method of Popperian methodology—called *alpha-beta method*—is developed for that purpose.

Chapter 2: Empirical Regularities of Capitalism: Eight Basic Facts

Because this book seeks to explain production and distribution in the capitalist system, the empirical regularities refer to both production and distribution data in the capitalist countries of the world. Moreover, because the book is interested in explaining not only within-country inequality, but also between-country inequality, a distinction is made between rich countries, called the First World (FW), and poor countries, called the Third World (TW).

Does history matter in the determination of the rich and poor status of capitalist countries? In order to answer this question, another distinction is introduced. Depending on their colonial history, two types of Third World countries are considered: those that have significant colonial legacy and those that do not. Therefore, the capitalist system will consist of three types of countries.

From the available empirical studies, it is possible to construct a set of empirical regularities, which is composed of eight facts. Seven correspond to production and distribution features in the post-World War II period, whereas the remaining one refers to the very long run relation between the process of economic growth and the environmental degradation. The initial periods of capitalist development are usually marked at the beginning of the industrial revolution, almost two hundred years ago.

For easy reference to this introduction, these regularities can be summarized as follows:

Fact 1: Existence and persistence of unemployment in the FW.

Fact 2: Existence and persistence of unemployment and underemployment (low-income self-employment) in the TW.

Fact 3: Existence and persistence of income gaps between ethnic groups in the TW.

Fact 4: In the short run, existence of interplay between monetary variables (money supply, interest rate, exchange rate) and real variables (output, unemployment, real wages) in both FW & TW.

Fact 5: In the long run, output and real wages are positively correlated in both FW & TW.

Fact 6: Income level differences between FW and TW are persistent over time.

Fact 7: The degree of within-country inequality is, on average, higher in the TW compared to that of the FW and the difference is persistent over time.

Fact 8: In the very long run, the economic growth process is accompanied by degradation of the bio-physical environment.

Why does the capitalist system operate in this manner? This why-question is a scientific question. According to Popperian epistemology, the scientific answer to a why-question is a theory. Any theory that intends to explain capitalism must not be refuted by these facts; moreover, a single refutation is sufficient to reject a theory. In science, facts are the masters and theories the servants, not the other way around.

Chapter 3: Can Standard Economics Explain the Eight Facts?

Standard economics is presented in this chapter in its elementary but fundamental form. Standard economics constitutes the current paradigm of economic principles and of public policies. The most popular textbooks in university courses include Keynesian economics for the short run macroeconomics, dealing with unemployment and inflation problems, and neoclassical economics for the long run study of output growth and its distribution. The same economics textbooks are utilized around the world, which implies that economic and social reality is homogeneous across all countries of the world, as is the case with textbooks of physics. Yes, economics is taught just like physics!

This book shows that standard economics cannot explain the eight facts. Standard economics is able to explain some of them, but not all. The assumptions of standard economics are then wrong and that is the reason for its failure. This conclusion is not a matter of taste or ideology; it is a matter of following the rules of scientific knowledge.

The need to search for a new economic theory of capitalism is then justified. A new economic theory—called *unified theory of capitalism*—is presented in this book.

The unified theory departs from standard economics in different ways, but also retains several of the assumptions of standard economics. According to Popperian epistemology, scientific progress is the result of an evolutionary process in which new and better theories are constructed upon the experience and failure of old theories. Thus the unified theory is standing upon the shoulders of giants, for giants were those scientists who built standard economics. The identification problem in epistemology says that when a theory fails it is impossible to identify which assumptions of the theory are responsible for its failure; hence, the new theory cannot be derived logically from

the old, but must be invented. Thus a new set of assumptions is proposed here to replace the set of assumptions of standard economics.

The new theory contains the following assumptions. The first refers to the role of the initial conditions of society (its history) in the economic process. Certainly, countries started capitalist development differently. In particular, they started with different degrees of inequality in the distribution of assets among individuals. These include economic assets (land, physical capital, and human capital) and political assets (degree of citizenship). This is called the *initial inequality*. Then countries started with different *factor endowments*, the amount of capital per worker. Some countries were born under-populated and some overpopulated. The unified theory assumes that these two initial conditions—initial inequality and factor endowments—are the essential factors in the economic process. The initial inequality (particularly inequality in political assets) depends upon the colonial history of countries. The other assumption is that the capitalist system comprises three types of societies, which are defined by their initial conditions: high, medium, or low initial inequality combined with under-populated or overpopulated factor endowments.

Other assumptions are old. Market and democracy are the essential institutions of capitalism. People act guided by the motivation of self-interest, subject to the institutional norms. Private property of capital is part of the norms. Labor exchange follows the rules of market exchange in the labor market.

Therefore, the selection of the set of assumptions of the new theory, which includes old and new assumptions, is a matter of intuition, a step in the trial and error procedure to construct better theories. The set of assumptions of any economic theory is somewhat arbitrary. This is why empirical corroboration of the theory is one of the general rules of scientific knowledge that is derived from Popperian epistemology.

The rest of the book is divided into three parts. Part I (Chapters 4, 5 and 6) presents the analysis of production and distribution in the short run for the three types of capitalist countries that constitute the capitalist system. Following Popperian epistemology, those societies must be seen theoretically, that is, as abstract societies that resemble the real world countries. To make sure that they are understood as abstractions of reality, they have been given names of Greek letters: epsilon, omega, and sigma, and defined as follows:

- *Epsilon society*: unequal in the initial distribution of economic assets, but equal in the initial distribution of political entitlements (one class citizenship), and under-populated. Intends to resemble the First World countries.
- *Omega society*: unequal as in epsilon, but overpopulated. Intends to resemble the Third World countries with no European colonial legacy or weak legacy.
- *Sigma society*: unequal in the initial distribution of economic assets as well as in the distribution of political entitlements (first and second class citizens), and overpopulated. Intends to resemble the Third World countries with strong European colonial legacy.

The scientific question is, of course, whether these abstract societies resemble well the real world societies that they intend to represent. The book shows that indeed

the three partial theories of capitalism (epsilon, omega, and sigma) resemble well the corresponding group of countries in the sense that they are able to explain production and distribution in the three types of countries, taken separately. The partial theories explain Facts 1-4.

Part II (Chapters 7, 8, and 9) compares the particular features of these three types of capitalist countries upon three factors: social order, physical capital accumulation, and human capital accumulation. These are supposedly crucial factors in understanding the process of growth and distribution in the long run.

Part III (Chapters 10 to 14) seeks to explain the capitalist system as a whole. The three partial theories are valid theories in the sense that they are able to explain production and distribution in each type of capitalist society, taken separately. The question is whether they constitute a unified theory: Do we have a theory that can explain the capitalist system, taken as a whole? Valid partial theories do not necessarily generate a valid unified theory, as we know from the experience of physics. Quantum theory explains the behavior of the small subatomic bodies and relativity theory explains that of the large bodies; however, they are inconsistent to each other. Theoretical physics is thus in search for *the theory of everything*, that is, a unified theory.

The unified theory of capitalism that is presented in this book is logically consistent with the partial theories; moreover, the unified theory explains Facts 5-8. The eight empirical regularities altogether are then explained by the unified theory.

According to Popperian epistemology, because facts do not refute the empirical predictions of the unified theory, there is no reason to reject the theory, and it may be accepted as a valid theory of the capitalist system, although only provisionally, until new empirical regularities or superior theories appear. Consequently, the valid theory can be applied in the social choices about public policies, which will become science-based public policies.

The synopsis that follows includes the results of each chapter, and also those of the book. Thus the reader can know what the book says in its most elementary and substantial form about production and distribution in the capitalist system, taken separately and as a whole. Economics deals with the functioning of human societies. Any subdivision of the economic process is just made for the sake of analytical convenience.

Chapter 4: Explaining Production and Distribution in the First World Countries

The epsilon society is an abstract society that intends to resemble the First World countries. The primary assumptions to construct epsilon society include given institutional rules and organizations (markets and democracy, and firms), in which individuals act guided by the motivation of self-interest. In addition, the theory assumes that the initial distribution of economic assets among social groups is unequal, but the society is equal in the entitlement of political assets, which makes epsilon society

socially homogeneous; finally, the initial factor endowments of machines and men in this society are such that it is under-populated.

Under these assumptions, the outcome of production and distribution from the economic process is determined by the international prices, the initial factor endowments, and the initial inequality. The outcome of the short-run static economic process (in which the process of economic growth is ignored) predicts the following empirical regularities of First World countries: Facts 1, 4, and 5. Therefore, so far epsilon society resembles well the First World countries; that is, these facts do not refute the predictions of the epsilon theory.

Chapter 5: Production and distribution in the Third World Countries with no Colonial Legacy

The omega society is also an abstract society, which intends to resemble the Third World countries that have no colonial legacy. This society is different from epsilon in one assumption only: the initial factor endowments of society are such that it is overpopulated. Omega society is overpopulated, but it is still socially homogeneous society. Under these assumptions, the production and distribution outcome becomes determined by international prices, initial factor endowments, and initial inequality. The outcome of the static economic process predicts the following empirical regularities of Third World countries: Facts 2, 4, and 5. Hence, so far omega society resembles well the Third World countries that have no colonial legacy; that is, these facts do not refute the predictions of omega theory.

Chapter 6: Production and distribution in the Third World with Colonial Legacy

The sigma society is an abstract society that intends to be like the Third World countries that have a significant colonial legacy. This society differs in one initial condition from omega society: there is inequality in the initial distribution of both economic and political assets. Like omega, sigma is overpopulated, but it is a socially heterogeneous and hierarchical society; that is, it is multiethnic and multicultural society, and a hierarchical one. Sigma is not only a class society, but it is a society with first and second class citizens. This social structure originates in the legacy of colonial institutions.

Under these assumptions, the production and distribution outcome becomes determined by international prices, the initial factor endowments, and the initial inequality. The outcome of the static economic process predicts the following empirical regularities of Third World countries: Facts 2, 3, 4, and 5. Hence, so far sigma society resembles well the Third World countries that have colonial legacy; that is, these facts do not refute the predictions of sigma theory.

To sum up Part I, epsilon, omega, and sigma theories are able to explain the short run process of production and distribution of each type of capitalist society. They are partial theories that can explain the components of the capitalist system, but taken

separately. According to the partial theories, income inequality is the common feature of the three types of capitalist societies. Another common feature is that the degree of income inequality depends upon the degree of the initial inequality in the distribution of economic and political assets.

Chapter 7: What is the Relationship between Social Order in Society and Its Degree of Inequality?

The question now is whether any degree of income inequality could be socially tolerated. A theory that assumes the existence of individual thresholds of tolerance to inequality is developed in this chapter. The empirical prediction of this theory is that the higher the degree of inequality, the higher the degree of social disorder in society will be. The cost of inequality for society is social disorder because income inequality constitutes the most fundamental source of social conflict in capitalist societies. The findings of the empirical international literature show consistency with this prediction: indeed more unequal countries tend to show higher degrees of social disorder, measured by crime rates, informality, and political instability. The theory of limited tolerance to inequality is not refuted by available facts. An implication of this theory is that highly unequal societies have lower quality of life.

Chapter 8: How Do Capitalists Choose the Countries in which to Invest?

Physical capital accumulation is assumed to be essential for the process of economic growth. Most technological progress is embodied in the new machines that firms buy and thus investment is the way to increase productivity continuously. Investment is considered the engine of economic growth.

What does the allocation of private investment to different types of capitalist countries depend upon? The theory presented here assumes that investors take into account the expected return and the risk involved in the investment projects; it also assumes that risk is higher where social disorder is higher, and the underlying degree of inequality is higher. Given the ambiguous effect of the expected return in the First World and the Third World, the theory predicts that private investment will flow mostly to the First World, in which the degree of inequality is lower and the risk factor is thus lower. The empirical evidence of the international literature is consistent with this prediction. Capitalist countries compete for private investment in international markets with their degree of equality, which is the underlying factor of social order.

Chapter 9: Is Education an Equalizing System?

Human capital is defined as the productive skills embodied in workers. Human capital is another essential factor for rapid economic growth. The theory presented here assumes that the human capital is fundamentally the outcome of the education process. It also assumes that the education process is socially hierarchical; namely, the output of human capital that emerges from the education system is higher for the social groups that are better endowed with economic and political assets. The theory predicts that the

education system in a capitalist society is not human capital equalizing. This result applies to the three types of capitalist societies, but it is stronger in sigma societies. The findings of the empirical international literature tend to be consistent with this prediction. Therefore, the initial inequality plays a significant role in determining the level and the distribution of the accumulation of human capital in a capitalist society.

Chapters 10&11: Why Do Income Gaps Between the First World and the Third World Persist in the Process of Economic Growth?

The unified theory of capitalism is presented in Chapter 10. Differences in the income levels and in the degree of inequality between the three types of countries are analyzed within the static model. This model predicts part of Facts 6-7.

A dynamic economic process is then constructed in Chapter 11 to analyze the determinants of economic growth. Each society has a growth frontier curve, which shows the growth path of output per worker in the long run. The dynamic model predicts that epsilon and sigma societies grow along different paths: the growth frontier curve is at higher level in epsilon compared to that of sigma; however, omega moves along a path that converges towards the growth frontier curve of epsilon. The ultimate factor that determines the separate growth frontier curves is the initial inequality differences. Given the relatively small size of omega type capitalist societies (Third World countries with no colonial legacy), the prediction of the model is consistent with Fact 6, which states that inequality between the First World and the Third World is persistent.

The dynamic model thus predicts that omega societies will naturally tend to become epsilon societies, but sigma societies will not. The prediction about omega societies also tends to be empirically consistent, although the number of capitalist countries that can be classified as omega is very small. In the history of capitalism, only Japan has succeeded in catching up with the First World, and there are few candidates (Taiwan and South Korea) that can do in the future (Maddison 1995). These three countries started capitalist development as omega societies, as most of the First World countries did, that is, as relatively equal and homogeneous societies. In contrast, most Third World countries started capitalist development as very unequal and hierarchical societies, as sigma societies, and are not able to catch up with the First World. The basic reasons are two: first, private investment flows mostly to more stable societies, that is, to more equal societies; second, there is no mechanism that can reduce the initial inequality in the process of economic growth.

On Fact 7, the unified theory indeed predicts that income inequality within epsilon, omega, or sigma societies will not tend to decline in the process of economic growth. The differences in the degree of within-country inequality will then tend to persist over time. The basic reason is that the income inequality of a capitalist society essentially depends on its initial inequality, which does not tend to decline naturally in the process of economic growth.

In the dynamic economic process of growth and distribution, therefore, there exists path dependence; that is, history matters. The initial inequality of countries is not erased naturally in the process of economic growth and thus between-country inequality

and within-country inequality will be persistent. The initial inequality, how capitalism was originated in today's capitalist countries, the colonial legacy of countries, is the ultimate factor that explains the observed persistence of inequality in the capitalist system.

Chapter 12: Could Higher Rates of Economic Growth Improve Social Progress?

One may argue that what is needed to solve social problems, such as inequality, is more rapid economic growth. In fact, this is the argument found in standard economics. But, are there limits to economic growth? The answer of standard economics is no; economic growth can continue forever. In order to analyze this question, the unified theory now includes into the economic process natural resources and the laws of thermodynamics, which deal with matter and energy relations.

According to the first law (the Law of Conservation), production of goods only rearranges matter and energy; that is, material inputs are transformed into material goods and waste, and free energy into used energy. According to the second law (the Entropy Law), production of goods transforms free energy into bound energy in the atmosphere, which degrades the bio-physical environment continuously and irrevocably. Since qualitative changes take place in the economic process, the mechanical dynamic model is now transformed into an evolutionary model. The evolutionary model predicts that there are limits to economic growth because production of goods implies a gradual reduction of the production potential of society, via depletion of non-renewable natural resources and pollution of the environment.

Capitalist societies have engaged in economic growth for almost two centuries, which has degraded the environment rapidly, which in turn has changed the carrying capacity of the environment to support human life, as we know it. This process cannot continue forever and qualitative changes in human life will take place at some finite time. Therefore, human societies will evolve qualitatively in some new directions, with new rules and organizations so as to adapt to the new physical environment.

The unified theory thus explains Fact 8: economic growth is accompanied by a degradation of the physical environment. Therefore, there is no such thing as sustainable economic growth. According to the unified theory, the process of economic growth and distribution in the very long run is not mechanical, but rather evolutionary.

In sum, the three partial theories epsilon, omega, and sigma are able to explain the functioning of the three types of capitalist countries, taken separately (the First World, the Third World with weak colonial legacy, and the Third World with strong colonial legacy); the unified theory is in turn able to explain the functioning of the capitalist system, taken as a whole. The reason is that none of the eight known empirical regularities of capitalism refutes the predictions of the unified theory. Following Popperian epistemology, we can say that there is no reason to reject the unified theory at this stage of our investigation; therefore, it can be accepted, although provisionally, until new empirical evidence is available or a superior theory is developed.

Chapter 13: What is New about the Unified Theory?

From Popperian epistemology we know that theory is a set of assumptions to construct abstract worlds in order to understand the real world. Therefore, the difference between the unified theory and standard economics, the current paradigm, lies in the assumptions. Also from Popperian epistemology we know that different theories predict different empirical relations. Hence theories can be evaluated not by their assumptions, but by the consistence between facts and the empirical predictions derived from the set of assumptions. These scientific rules constitute the criteria with which the new theory and standard economic are evaluated. The result is that the set of assumptions of the unified theory is able to generate a set of empirical predictions that is consistent with the eight basic empirical regularities of the capitalist system; by comparison, the set of assumptions of neoclassical theory and Keynesian theory generate empirical predictions that can explain some of those facts, but not all.

The unified theory is not only a new economic theory of capitalism. It also sets the foundations of new economics, in which the economic process takes place under environmental distress. Economics must deal with the fundamental problem of our time: the fate of the human species. This view may be called Modern Economics to distinguish it from Old Economics or Standard Economics.

Chapter 14: What Are the Public Policy Implications of the Unified Theory?

The basic causality relation established by the unified theory is that the three outcomes of the economic process—growth, inequality, and environment degradation—are explained by the initial inequality under which capitalist countries were born. According to the unified theory, and contrary to neoclassical theory, government policies are proximate factors; technological progress is also proximate factor, because these factors are determined by the initial inequality, that is, the power structure of society. Therefore, the ultimate factor explaining the economic process is the initial inequality of countries and the initial inequality of the capitalist system as a whole. The initial inequality is possibly not the only ultimate factor; but according to the unified theory, it is the essential factor. This causality relation has been proven to be consistent with the empirical regularities of capitalist development.

Once an economic theory has survived the falsification process and has become a valid theory, then, and only then, its public policy implications can become science-based policies. These implications refer to both social objectives and policy instruments. The unified theory is a valid theory and its policy implications can then be analyzed.

Any public policy reflects a social choice, no matter how imperfect the mechanism of social choice is. It is evident that the current public policy followed in most capitalist countries seeks to maximize economic growth. But according to the unified theory there is a trade-off between growth and social progress. Alternative public policies could seek to improve quality of life of present and future generations, which implies dethroning economic growth as the sole social objective. According to

the unified theory, the ultimate factor that explains the outcomes of the economic process in the capitalist system is the initial inequality, that is, the degree of concentration in economic and political power. Therefore, as long as this power structure remains unchanged, the outcome of the economic process will persist.

The alternative social objective to pro-growth would imply the redistribution of the current economic and political power in the capitalist system as a whole. Changing the current outcome of the economic process implies the application of public policies that seek to reduce the initial inequality, that is, to break with history; in particular, it implies citizenship equality at national and international levels. Thus institutional innovations would need to be introduced into the democratic system to make it stronger and thus a more representative social choice mechanism.

In light of the unified theory, public policy options can be summarized as follows: Either we maintain the current pro-growth policies, which tends to maintain or even increase the high degree of inequality—and the consequent social disorder—and to increase the rate of environmental degradation, or we use our scientific knowledge about the economic process to create a social world in which the quality of life is better for the present and future generations. Changing paradigms in economics is difficult. We are used to the idea of growth and a world of no-growth is thus unthinkable. But some theorists, old and new, have shown how such no-growth society would operate, and even reach higher quality of life.

The alternative public policies have never been applied. Why? By choosing in favor of pro-growth public policies, economic and political elites have revealed their preferences. They benefit more from this policy than from the alternative policies. According to the unified theory, the high degree of income inequality in the capitalist system does not diminish in the process of economic growth, which has empirical support. Hence, dethroning economic growth as the sole objective of public policy is against their interests. At the same time, standard economics (the current paradigm) also implies pro-growth policies. Enthroning alternative policies that seek capitalist societies with high degree of quality of life, therefore, requires changes in both the current economic paradigm and in the current concentration of economic and political power, at national and international levels. This is the fundamental problem of our time.

The book started with the idea of solving a problem: What are the determinants of social progress under capitalism? It concludes that the ultimate factor is the inequality in the distribution of economic and political assets. But it then ends with a new problem: What are the determinants of the economic and political power concentration under capitalism and how could they be changed? This progressive way of explaining the real world is in the nature of scientific knowledge.

About the Reach of the Book

The nature of the science of economics is shown along this book. Physics is considered the exemplar of science. But the late Harvard biologist Ernst Mayr stated that “biology is, like physics, a science, but biology is not a science like physics” (Mayr 1997, p.32). Paraphrasing Mayr, the book will show that economics is, like physics, a science, but

economics is not a science like physics. Economics is more like biology, a historical and evolutionary science.

The ontological universalism of physics cannot be applied to economics; on the contrary, the existence of different societies under capitalism calls for partial theories and thus for the need of a unified theory to explain the whole and the parts of the capitalist system. The unified theory of capitalism that will be presented in this book should then be seen as a modest contribution to that goal.

This is a book about the scientific explanation of the production and distribution process in the capitalist world. A new theory will be presented, which is able to explain this world. This book is not about a new version of standard economics or its derived doctrines that are dominant today; on the contrary, the book intends to challenge standard economics. But this objective will be pursued in the realm of science, by using the rules of scientific knowledge. The current degree of environmental degradation has meant for economics a new context in which to study the economic process in our time, with new foundations. Thus the book seeks to present the new foundations of the modern science of economics.

The book requires basic knowledge of standard economic theories. This is just in accord with the principle that scientific progress is a cumulative process, in which new contributions are standing on the shoulders of giants. Hence, and hopefully, the reader of this book not only will be able to learn a new economic theory, but also will understand better standard economic theories. Basic knowledge of mathematics is also needed. But the reader should think economics—not mathematics.

CHAPTER 1

THE RULES OF SCIENTIFIC KNOWLEDGE

Why has the progress of scientific knowledge in the social science proceeded at a pace that is slower than that of the natural sciences? A possible reason is the limitation of data, in quantity and quality, in the social sciences; in addition, the instruments of measurement for the social phenomena are very imperfect. The other reason seems to rest upon the limited use of methodology in the construction of scientific knowledge in the social sciences. Compared to the natural sciences, the social sciences seek to explain the functioning of the social world, which is a much more complex world than the physical world; therefore, understanding the social world is more demanding on methodology than understanding the physical world.

Economist Paul Samuelson wrote in his classic book *Foundations of Economic Analysis* about this comparison as follows:

[This] book may hold some interest for the reader who is curious about the methodology of the social sciences...[I]n a hard, exact science [as physics] a practitioner does not really have to know much about methodology. Indeed, even if he is a definitely misguided methodologist, the subject itself has a self-cleansing property which renders harmless his aberrations. By contrast, a scholar in economics who is fundamentally confused concerning [methodology] may spend a lifetime shadow-boxing with reality. In a sense, therefore, in order to earn his daily bread as a fruitful contributor to knowledge, the practitioner of an intermediately hard science like economics must come to terms with methodological problems (Samuelson 1947, pp. viii-ix).

The separation between the natural sciences and the social sciences made by Samuelson between hard and intermediately hard sciences will be transformed into complex and less complex sciences in this book.

Methodology is another name for epistemology (from the Greek *episteme*, knowledge). Epistemology deals with the logic of scientific knowledge from which a practical set of rules to arrive at scientific knowledge can be derived. Economics is ambiguous upon epistemology. A practical set of scientific rules for economics can be derived from the epistemology of Karl Popper (1968). This is done in this chapter.

1.1 Scientific Rules from Popperian Epistemology

Scientific knowledge seeks to establish relations among objects. The objects can be mental or physical. Formal sciences study the relations among mental objects, whereas

factual sciences study the relations among material objects. Mathematics and logic are examples of formal science; physics and economics are instances of factual sciences.

Scientific knowledge takes the form of a proposition that intends to be error free. Scientific knowledge is therefore a particular type of human knowledge.

What would be the criterion to accept or reject a proposition as scientific? It depends upon the type of science. In the formal sciences, the criterion seems to be rather straightforward: the relations established must be free of internal logical contradictions. In the factual sciences, by contrast, the criteria are more involved. The propositions of a factual science must be free of internal logical contradictions as well. However, this criterion constitutes just a necessary condition; empirical consistency or empirical testing between the propositions and the facts will also be required.

According to the epistemology developed by Karl Popper, scientific knowledge that seeks to explain reality cannot be attained by using inductive logic. There is no such thing as inductive logic; that is, there is no logical way to go from particular empirical observations to general laws. His classical example is: “No matter how many instances of white swans we may have observed, this does not justify the conclusion that “all swans are white.”

Scientific knowledge can be attained by using deductive logic. The logic is as follows: theory is constructed to explain reality; from theory some conclusions about reality are derived by logical deduction. These conclusions, which will refer to propositions about reality, are then submitted to confrontation against reality. If the empirical predictions and reality are consistent, we have no reason to discard the theory; but if they are inconsistent, then the theory has been proven false and it has been falsified. Hence, a theory must generate falsifiable empirical propositions.

An empirical proposition is falsifiable if in principle it can be false. This is the principle of falsification. The proposition “It will rain or not rain here tomorrow” is not falsifiable, for it is a tautology, whereas the proposition “It will rain here tomorrow” is. Falsification is therefore the demarcation principle between scientific and non-scientific propositions.

The basic scientific rules that can be derived from Popperian epistemology include:

- (a) Scientific theory is required to explain the real world. No theory, no explanation.
- (b) Falsification is the criterion of demarcation. The scientific theory must be falsifiable in the sense that the empirical propositions derived from it, by deductive logic, must be falsifiable.
- (c) If its empirical predictions are refuted by the reality, the scientific theory is rejected; if they are not, the theory is accepted provisionally until new data or superior scientific theory appears.

The question now is to see whether these scientific rules can be applied to economics. In order to answer this question, we must firstly know the nature of the scope of economics.

1.2 The Economic Process

Economics is a social science. It seeks to explain the functioning of human societies in regard to a particular aspect: the determinants of the production of goods and its distribution between social groups. The type of social relations that economics studies refer to the relations between people regarding production and distribution of goods; goods are the cement that link people. This is the scope of economics.

Human societies constitute complex realities. The notion of complexity refers to the large number and the heterogeneity of the elements that constitute the reality, and to the multiple factors that shape the relations between those elements. Human diversity together with the multiplicity of human interactions makes human societies intricate realities. The simple fact that individuals in a human society are not identical, as compared to homogeneity of atoms in the physical world, suggests that the social world is more complex than the physical world. Human societies are complex systems of interacting individuals in which individuals themselves are complex systems.

How can a complex social reality be subject to scientific knowledge? The first assumption is that the complex social reality can be transformed into a process, in which the social relations occur regularly and repeatedly. The second is that the complex social reality is reducible to a simpler abstract world by setting aside elements of the process that are not important to understand the social world. This is the well-known method of abstraction.

Conceptually, a process is a series of activities carried out in parts of the real world, having a given duration and having a purpose, and being repeated period after period. The essential characteristics of a process include the existence of a boundary that separates the outside from the inside because the process refers to a partial aspect of social reality; the other is repetition: process always implies repetition. On the structure of the process, there are elements that cross the boundary from outside the process—called the *exogenous* elements—and those that cross the boundary from inside the process—the *endogenous* elements. In a process, there is also an underlying mechanism by which the exogenous elements are transformed into the endogenous elements (Georgescu-Roegen 1971, Chapter IX).

The transformation of a complex social reality into a process implies the separation of all elements of reality into endogenous and exogenous. The complete list of endogenous and exogenous elements of a process would include observable and non-observable elements. Call the observable elements *endogenous variables* and *exogenous variables*. The method of abstraction must now be applied to reduce the process to a simpler form, which is called *process analysis*. The use of abstraction implies the selection of the most significant endogenous and exogenous variables together with the most significant mechanisms of the process. Certainly, to present the complete list of the variables of a process would be equivalent to constructing a map to the scale 1:1. As in the case of the map, a complex reality cannot be understood at this scale of representation. Abstraction implies that some variables and some mechanisms must be ignored. This is how a real world is transformed into an abstract process, into an abstract world, in which only the supposedly important variables and mechanisms are included and the rest are just ignored.

The analytical representation of process analysis is illustrated in Figure 1.1. The segment t_0 - t_1 represents the duration of the process, which is going to be repeated period after period. X is the set of exogenous variables and Y is the set of endogenous variables. The shaded area indicates the underlying mechanism by which X and Y are connected.

What happens inside the process is not observable, as indicated by the shaded area in the figure. If it were, the interior of the process would be considered another process in itself, with other endogenous and exogenous elements and another mechanism; the latter mechanism would also be observable and then constitute another process, and so on. Thus, we would arrive to the logical problem of an infinite regress. (Science avoids this trap by making assumptions about the initial conditions of a process.) Ultimately, there must be something hidden beneath the things we observe. Science seeks to unravel those underlying elements.

Because a process repeats itself period after period, the relationships between exogenous and endogenous variables could be observed continuously, and could then develop systematic relationships or empirical regularities among them. The existence of empirical regularities is a necessary condition for scientific knowledge. A chaotic world—where regularities are absent—is much harder to understand; therefore, it will not be part of process analysis.

How do we decide which elements are important in a process and which are not? How is an abstract world or abstract process constructed? They are established by a set of assumptions, that is, by constructing a *scientific theory*.

The endogenous variables constitute the object of study, the phenomena that the theory intends to explain. The exogenous variables are the explanatory factors. The construction of an abstract process is made through the introduction of assumptions about which exogenous variables of the process are important and which are not. In addition, assumption must be made about what are the relevant underlying mechanisms of the process. This set of assumptions constitutes a scientific theory. Hence, the theory determines the endogenous and exogenous variables and the particular mechanisms of the abstract world. A theory is, therefore, a logical artifice by which a complex real world is transformed into a simple abstract world. This abstract world will be much simpler to understand.

A method to transform the real social world into an abstract world is therefore needed. This method must be consistent with Popperian epistemology and with process analysis. The alpha-beta method constitutes such a method, which is presented now.

1.3 The Alpha-Beta Method

The method starts with the following definition:

In terms of the logical ordering of its propositions, a scientific method can be seen as a set of alpha and beta propositions, in which beta propositions are logically derived from the alpha and no alpha proposition is derived from

another alpha; hence, alpha propositions constitute the primary assumptions of the theory (adapted from Georgescu-Roegen, 1971, p.26).

Alpha is the set of primary assumptions of the theory that attempts to explain reality. (The term “primary” leaves open the possibility of “auxiliary” assumptions, as will be shown later on.). As indicated above, the assumptions refer to the components of a process: the supposedly significant endogenous and exogenous variables and the supposedly significant underlying mechanisms that connect them. With these assumptions, an abstract economic process can be constructed.

Can the assumptions of a theory be established from empirical observations? No, they cannot. The reason is that the theory seeks precisely to explain those observations. What we can get from reality by empirical observation is a description of it, not an abstraction. The listing of all elements of the process by itself cannot discover the essential and non-essential elements.

Do the assumptions of a theory need justification? No, they do not. The assumptions are of the nature of axioms, in the sense that they need no justification. If the set of assumptions needed justification, another set of assumptions to justify them would be needed, which in turn would need another set to justify the latter, and so on; hence, we would reach the logical problem of infinite regress.

The assumptions of a theory are non-observable propositions (such as about the motivations of social actors), which may be empirically false because they are not universally true. They do not intend to be empirically true, but just a good approximation of reality. The assumptions do not imply that the factors left out do not exist; they only say that even if other factors do exist and affect the outcome of the process, those effects are not significant.

The set of assumptions of a theory are established arbitrarily. However, according to Popperian epistemology, a theory is constructed as part of an algorithm, of a trial and error process, the aim of which is to reach a good theory. A good theory is the one that has constructed a simple abstract world that resembles well the real world. If a theory fails, a new set of assumptions is established to form a new theory, a new abstract world is thus constructed; if this second abstract world does not resemble the real world, the theory fails and is abandoned, and a new set of assumptions is established, and so on. The best theory is found by an evolutionary process in which we assist to the funeral of some theories. An implication of the falsification principle is that theories must be mortal. Hence, what the assumptions of a theory need is not justification, but falsification against data of the real world, independently of how or why the assumptions were chosen.

How is this falsification made? This is the role of beta propositions. Beta propositions are the result of the logical derivation from the set of assumptions. The logical derivation is about the relations between the endogenous and exogenous variables of the theory. Therefore, three characteristics of beta propositions can be established:

- Beta propositions constitute the empirical predictions of the theory; that is, the relationships between the endogenous and exogenous variables of the theory.

- Beta propositions show *causality* relations: the effect of exogenous variables upon endogenous variables. Note that causality requires a theory; that is, no theory, no causality. The theory seeks to explain the changes in the endogenous variables in terms of changes of the exogenous variables, given the underlying mechanisms.
- Beta propositions are *falsifiable*. The relations between exogenous and endogenous variables that beta propositions derive from the theory are observable and hence make the theory refutable.

On the last characteristic of falsifiability, what the beta propositions say may or may not coincide with what we observe in the real world. Although a beta proposition is logically correct, it may be empirically false. The reason is that the set of assumptions contained in the alpha propositions were selected arbitrarily, for there is no other way to select them. From the way beta propositions are derived, it follows that for each endogenous variable there will exist a beta proposition; hence, there will be as many beta propositions as there are endogenous variables in the theory.

A good theory must therefore be falsifiable through its beta propositions; that is, in principle, a good theory must be mortal. “Men die when God so wishes” is an example of an unfalsifiable proposition. The reason is simple: God’s wishes are unobservable. Shamans are particularly good at using unfalsifiable propositions to deal with sick people: “If you have faith on my medicine you will get well.” If the person complains that he is not getting well, the shaman could say, “You had no faith in my medicine.” This statement never fails because faith is unobservable. Apart from unobservable elements, there is another case in which a proposition is unfalsifiable: when it predicts all possible outcomes (“It will rain or not rain here tomorrow”). These types of propositions are tautological, useless for scientific knowledge. These examples illustrate the principle that a “theory” that cannot generate beta propositions becomes immortal and is actually not a theory.

If, in spite of the abstraction, the simple abstract world resembles well the complex real world, the theory is then a good theory. The abstract world resembles well the real world; accordingly, we say the theory explains the reality. This good theory can now be called the *valid theory*. A valid theory is the one that gives rise to beta propositions that have not been refuted by facts, although in principle they could have. This is similar to the idea that an honest person is one who having had the opportunity of committing a crime did not do it; but if he never had the chance, we cannot say.

How do we decide to accept or reject a theory? If all beta propositions coincide with reality, the theory is not refuted by the available facts; if at least one beta proposition fails, the theory fails to explain the reality. Beta propositions represent the necessary conditions to accept the theory as valid; that is, if a single beta proposition fails, the theory necessarily fails because the beta proposition was logically derived from the set of assumptions of the theory.

The last property can be illustrated with a simple example. Consider a theory that states that Figure F is a square (suppose the Figure F is unobservable by the researcher). By definition a square is a rectangle with all four sides equal. If this is taken as the assumptions of the theory, then the following beta proposition can be logically derived: the two diagonals must be equal. If the available empirical evidence shows that

the diagonals are not equal, Figure F cannot be a square. The theory is refuted by this fact. However, if empirically the diagonals are equal, we cannot conclude that F is a square (it could be a rectangle). Truth is much harder to establish in science.

As this example shows, a theory can only be *corroborated* by facts; it cannot be *verified*, proven to be true. The latter implies consistency with fact by necessary and sufficient conditions, the former by necessary conditions only.

Which particular assumptions are responsible for the failure of a theory? As we have seen, theory is a logical artifice that allows us to arrive to scientific knowledge. Theory constitutes an abstraction of reality. If there is no theory, there is no scientific knowledge. The theory needs empirical confrontation against reality. The prior set of assumptions needs posterior validity. The reason is that the assumptions were arbitrarily established. If in this confrontation theory and reality are inconsistent, theory fails, not reality; that is, the arbitrary selection of the *set* of assumptions is proved to be wrong. Thus the particular assumption responsible for the failure is not identifiable. Due to this problem of identification, a new theory could not be derived logically from the old one, but it would have to be invented, using a new set of assumptions.

To understand a complex world the use of theory is a necessity. As shown above, theory is an abstraction of the real world, in which only some elements of the process are taken into account. The beta propositions constitute the empirical predictions of the theory and are utilized to seek refutation of the theory. Logically, therefore, a beta proposition can fit only the general or typical cases of the real world observations. Due to the use of abstraction, it may not fit all the observed cases. Therefore, the refutation of a theory needs to be based on statistical analysis.

The empirical prediction of a theory must then be confronted against the relationship between the average values of the observed variables. If a theory on income inequality predicts that incomes of workers depend directly upon their years of schooling, this prediction is refuted by facts if the *average* incomes are inversely related to *average* years of schooling; it is not refuted if the relationship is direct.

A single empirical observation that contradicts a beta proposition is insufficient to refute a theory, for the statistical value of one observation is nil. If one worker has a high number of years of schooling but a low income, that observation cannot refute the theory. That observation could just correspond to an exception, a deviation from the average due to other factors or purely by chance, which is called statistical error. A single counter-example is sufficient to invalidate a theorem in mathematics, but it is not sufficient to refute a theory. Accordingly, a distinction must be made between *error* of a theory (statistical error) and *failure* of a theory.

The assertion that a theory explains reality therefore has a precise meaning: its beta propositions are statistically *consistent* with empirical data. On the other hand, if empirical data do not fit the beta propositions, the theory is simply *false*. In this case, a new theory should be formulated and the same algorithm continued.

The continuous interaction between theory and empirical data is the cornerstone of the alpha-beta method. The logic of this method is thus clear. A complex reality is reduced to an abstract process—where there is repetition, and where regularities in empirical relationships can occur—with the use of a theory (alpha propositions); from

the theory, empirical predictions are logically derived (beta propositions), which are falsifiable; the falsification is made through statistical analysis. If the beta propositions are consistent with facts, there is no reason to reject the theory, and it is accepted as valid; if the beta propositions are refuted by facts, then the theory is rejected and, thus, a new theory is constructed. Scientific knowledge is achieved through this algorithm.

The alpha-beta method is represented diagrammatically in Figure 1.2. From the set of alpha propositions α_1 , the set of beta propositions β_1 is logically derived (indicated by the double arrow). The set β_1 must then be subject to statistical testing (indicated by the single arrow). While the double arrow indicates deductive logic, the single arrow indicates operational procedure, or the task to be performed. Statistical testing of the theory (indicated by the symbol \approx) implies seeking a statistical consistency between beta propositions and the available set of statistical relations or associations between endogenous and exogenous variables, the dataset b . If statistically $\beta_1=b$, then α_1 is consistent with reality, there is no reason to reject the theory, and then we may accept it; if statistically $\beta_1 \neq b$, then reality refutes the theory α_1 , and another theory α_2 should be developed; thus the algorithm is continued.

An example of the Popperian falsification principle using the alpha-beta method is the following, taken from biology:

α : Plants seek to maximize solar energy.

β : Then, plants will position their leaves in a particular distribution so as to maximize exposure to sun: each leaf collects its share of sun without interfering with other leaves.

b : We observe that the leaves of trees form canopies that that type of distribution.

The first proposition is one of the alpha-type: it is an assumption and is unobservable. The underlying mechanism is the photosynthesis. The second is the beta proposition derived from alpha by using deductive logic. If the first is true, the second follows necessarily. The third statement is about what the data say. Because the beta proposition coincides with empirical data b , the theory cannot be rejected. Fact b does not refute the theory. Moreover, the theory explains why the leaves of trees form this type of canopies, that is, the why-question is solved.

In the case of falsifying several theories at the same time, given data set b , some theories will be false and some will be consistent. Those theories that survive the entire process of falsification will become the valid theories. They will reign until new dataset, new statistical testing methods, or a superior theory appears. A theory is superior to the others if it generates the same beta propositions as the others, but in addition generates other beta propositions that are consistent with facts, which the other theories cannot. A theory is thus superior to others when it can explain the same facts that the others can and also some additional facts that the others cannot.

It is clear from the alpha-beta method that data alone cannot explain real phenomena. Data alone can produce statistical association or correlation between empirically defined variables. But there is no logical route from statistical association or correlation to causality (no matter how sophisticated the statistical technique is). The

logical route can go from alpha-beta propositions to dataset b, but not vice versa. Causality requires an underlying theory because exogenous and endogenous variables can only come from a theory. Reality can be understood only in the light of a theory. The common argument “Let the empirical data speak for themselves” is not epistemologically justified. To recall Darwin’s dictum, *all observations must be for or against a theory*.

In sum, the alpha-beta method is in accord with Popperian epistemology. The rules of the alpha-beta method are consistent with the three rules of Popperian epistemology shown above: (a) The rule that theory is needed for explanation is given by the alpha proposition; (b) The rule of falsification is given by the beta proposition; (c) The rule of rejection-acceptance of a theory is given by the consistency between the beta proposition and the empirical data. Therefore, the alpha-beta method is a particular method of Popperian methodology. Moreover, while Popperian epistemology gives us only general rules of scientific knowledge, the alpha-beta method makes them more operational.

1.4 The Alpha-Beta Method in Economics

The scope of economics can now be defined in terms of process analysis. Figure 1.1 can be used for this purpose. Economics is a social science that studies the economic process, which is defined as the process of production of goods and the distribution of those goods among social groups in human societies. Production and distribution are the endogenous variables of any economic theory. The factors that constitute the ultimate cause of changes in production and distribution are the exogenous variables and the underlying mechanisms are particular to theories. An economic theory contains a set of primary assumptions from which beta propositions are derived. There are however some particularities in the application of the alpha-beta method to economics, which are spelled out now.

Economic Theories Include Universal and Society-specific Alpha Propositions

Physics is considered the exemplar of factual sciences. A characteristic of physics is that it assumes ontological universalism, that is, the physical world is seen as a single world. Therefore, the theories of physics are supposed to be valid for any place on Earth.

As a social science, economics assumes that societies do not conform to an ontological universalism. Then economics could not generate conclusions that are valid for every human society, independent of place and time. The other extreme view, that economics is valid only for specific societies only, cannot be assumed either because economics cannot explain individual cases, only aggregates. The assumption made by economics about its scope lies in between. Economics has two types of assumptions on the economic process: one is general, common to all human societies, and the other is specific to particular types of societies. The universal assumptions are presented now, whereas the specific assumptions will refer to capitalist societies, and will be presented in the rest of the book.

Economics assumes that the nature of the production process is the same in all societies. The production of goods requires a stock of primary factors of production (labor, capital, land and other natural resources). Given the technological knowledge of society, which refers to the knowledge on how to produce goods, it is assumed that the quantity of total output depends on the quantity of primary factors of production used in the production process. This relationship is called the *production function*. An important concept is derived from the production function: the *production frontier* of society. This is the boundary that separates the menu of goods that society can produce with its technology and factor endowments from those that cannot.

Another universal assumption is that human necessities are unlimited. The nature of the economic problem faced by any human society can now be logically derived from these two assumptions. The assumption of the existence of the production frontier implies that, on one hand, society has a limited capacity to produce goods with its factor endowments and technology; while on the other hand, the quantity of goods desired in society is unlimited due to the nature of human necessities.

In order to solve the economic problem, societies establish the rules under which the economic process must operate. Economics assumes that societies operate with given institutions, which include both social organizations and the social norms of the economic game (North, 1990). As in a football game, the economic process needs organizations, that is, social actors that carry out the economic activity; it also needs a set of rules under which these agents will interact. Economic theories must then make assumptions about what the essential components of institutions in society are.

Another universal assumption is that social actors have an economic rationality. The logic underlying the behavior of people is called rationality. This logic refers to the motivations that guide the actions of people. Economics assumes that individuals are rational in the sense that they act guided by particular motivations and aims, which are shaped by the social norms of the society in which they live; that is, they act adequately or appropriately or in accordance with the institutional context (Popper 1985). The rationality assumption implies that people confronting the same institutional context will tend to behave similarly.

Finally, there is the universal assumption about the weight of the initial conditions of a particular type of society. This weight is significant, which implies that history matters in understanding the economic process.

The universal alpha propositions can be summarized as follows:

α_0 (1). *The scarcity postulate*: Human societies face the economic problem: the maximum flow of goods that society can produce is limited, whereas the quantity of goods desired in society is unlimited. Nirvana societies do not exist or are the exception.

α_0 (2). *The institutional postulate*: In order to solve the economic problem, societies seek to establish a particular institution, that is, a set of rules and organizations, which regulate the activities of production and distribution.

α_0 (3). *The rationality postulate*: Individuals act guided by motivations, which are shaped by the institutional context of the society in which they live.

α_0 (4). *The initial conditions postulate*: Types of human societies differ in their origin (history), which is characterized by two initial conditions: the factor endowments and the distribution of economic and social assets among individuals. The first refers to the aggregate factor endowments, which determine labor productivity levels in society; the second refers to the initial degree of inequality in society.

An economic theory that seeks to explain production and distribution in a particular type of human society will still consist of alpha and beta propositions. But the alpha propositions will have two components: the universal assumptions—presented above and labeled with the zero subscript—and the society-specific assumptions, which will have a subscript to identify the particular type of society under study. If economics pursues to be a social science that studies the process of production and distribution in human societies, the society-specific assumptions must be logically consistent with the universal assumptions presented above.

An Economic Theory Is a Family of Models

An economic theory cannot be falsified if it contains only general alpha propositions, which are in the domain of human intuitions, because beta propositions could hardly be derive from them. If this is the case, additional assumptions, called *auxiliary assumptions*, are necessary to make the theory operational and thus to derive beta propositions. The new set of assumptions constitutes a *model of the theory*. A theoretical model then includes two subsets of assumptions: the alpha propositions, which are the primary assumptions, and the auxiliary assumptions. These two subsets of assumptions must be logically consistent to each other; that is, the auxiliary assumptions cannot contradict the primary assumptions.

An economic theory can then be seen as a family of finite models. The set of alpha propositions constitutes the common element or the core of the family. A model is also a logical system, that is, a system free of internal logical inconsistencies. Beta propositions will now be derived from models. The theory will now be subject to the process of falsifiability through its models.

The Algorithm of the alpha-beta method in economics is shown in Figure 1.3. Given a theory α , a set of consistent auxiliary assumption A' is included to construct the model α' , from which the set of empirical predictions β' is derived, which is tested out against the set of empirical data b . If the theory does not fail, it can be accepted; if it does fail, then α' model fails, but not the theory. Using auxiliary assumption A'' , another model (α'') is constructed and submitted to the falsification process; if it fails, then another model is constructed, and so on, until model α^n . If the n models fail, then the theory fails and a new theory is constructed, and the algorithm is applied again.

The alpha-beta method applied to economics can also be illustrated in Figure 1.2. Let α_1 and α_2 be now two models of a theory called α theory. The falsification algorithm is now applied to the models. If the beta propositions of the first model are not refuted by empirical data, the model can be accepted, and then the theory can also be accepted; if they are refuted by empirical data, the model is rejected, but the theory is not. The second model will then be submitted to a test, and so on. For the theory to be

refuted, all models of the family must fail. This algorithm requires that the number of models be finite; that is, it requires that the theory can generate only a limited number of possible auxiliary assumptions. If all models fail, the theory fails, and a new theory is needed.

The number of models of a theory must be finite if the theory is going to be falsifiable. How to justify it? The criteria utilized to construct models refer to the use of auxiliary assumptions. But what these assumptions are about? Basically, auxiliary assumptions refer to the particular context in which social actors operate. In the study of the capitalist society, for example, consider an economic theory that assumes the market system and the democratic system as the basic institutions. Then the following possible social contexts can be established:

- (a) On the market system: assume the particular degree of market power (perfect competition—absence of market power—, monopoly, or oligopolistic competition);
- (b) On the democratic system: assume the particular political power structure (highly participatory, representative, or authoritarian);
- (c) On the nature of the economic process: static, dynamic, or evolutionary)

A combination that results from the selection of one element from each category will determine a particular model of this economic theory of capitalism. The total number of possible combinations is thus finite in this example.

It should be noted that the use of auxiliary assumptions and models do not constitute protective belts of theories to avoid falsification as some epistemologists have suggested (the so-called Duhem-Quine problem.) On the contrary, the use of finite number of models makes an economic theory falsifiable.

The Existence of Equilibrium is Fundamental for Falsification

The outcome of the economic process is a particular value of production and distribution, which implies that these values will be repeated period after period. However, there are different logical ways in which the economic process is repeated. In order to determine those forms, a definition of equilibrium in the economic process is needed. This can be stated as follows:

An economic process is said to be in an *equilibrium* situation, if, and only if, no social actor has both the power and the incentive to change the outcome of production and distribution.

The conditions of equilibrium may be observable or unobservable. In the former case, the values of the endogenous variables are restricted to take values in a given range (unemployment rate equal to zero, for example) under any equilibrium situation. The observable conditions will therefore be logically derived from the model of a theory and will take the form of an empirical prediction of the model. For simplicity, we could call it beta proposition as well, although it does not refer to causality relations in this case.

In the economic process, the notion of equilibrium may refer to a static process or a dynamic process. In the first case, the model assumes that the relationships among endogenous variables are contemporaneous, whereas in the dynamic process some or all are inter-temporal.

Firstly, consider a static process. *Static equilibrium* implies the repetition of the same values of the endogenous variables, period after period, as long as the values of the exogenous variables remain constant. Static equilibrium may be stable or unstable. It is stable when the value of the endogenous variable spontaneously restores its equilibrium position whenever it falls out of equilibrium (the classical metaphor is a ball inside a bowl); otherwise, the equilibrium is unstable (a ball on top of bowl that is placed upside down). Certainly changes in exogenous variables will imply changes in the equilibrium values of the endogenous variables. Beta propositions are determined by using the method of *comparative statics*. The method of comparative statics shows the changes in a system from one position of equilibrium to another without regard to the transitional process involved in the adjustment (Samuelson, 1947, p. 8). This method requires stable equilibrium.

In a dynamic process, given the inter-temporal relationships among endogenous variables; and given the initial conditions and the values of the exogenous variables, the equilibrium values of the endogenous variables of today determine their equilibrium values of tomorrow, which in turn determine their equilibrium values of the following period, and so on. Endogenous variables change just with the passage of time. Thus, *dynamic equilibrium* implies a particular trajectory over time of the endogenous variables, as long as the values of the exogenous variables remain fixed.

Under certain conditions, the dynamic equilibrium is just the sequence of static equilibrium situations. Therefore, if the static equilibrium is stable, the dynamic equilibrium is too: a situation that is for some reason out of the equilibrium trajectory will move spontaneously to the equilibrium trajectory, which is called *transitional dynamics*. In the dynamic process, beta propositions are determined by using the method of *comparative dynamics*, which shows the effect of changes in the exogenous variables upon changes in the equilibrium trajectory, including the transitional dynamics involved in the adjustment.

The economic process can take the form of either static or dynamic, depending on the assumptions of the theoretical model. Hence, when applied to economics, the representation of a process, as shown in Figure 1.1, can be understood either as a static process or as a dynamic process. In the first case, Y will change if, and only if, X changes; in the dynamic case, Y will move along a given trajectory over time if X remains fixed and will shift to another trajectory if, and only if, X changes.

Static or dynamic equilibrium assumes that the economic process is a mechanical one. It can be repeated forever. If the economic process is viewed differently, as subject to qualitative changes as it is repeated, then we have an *evolutionary economic process*. In the process of production and distribution, societies may change qualitatively, which implies that the economic process cannot be repeated forever and its explanation requires evolutionary theories.

It should be clear from the definition of equilibrium that it has no normative implications. Equilibrium could be socially desirable or undesirable. It does not show

social optimality in any sense. Equilibrium does not mean absence of social conflict either; it does not imply social harmony. Equilibrium is just the aggregate outcome of social interactions, in which no social actor has the power and the will to change the situation.

How does falsifiability operate in mechanical and evolutionary processes? Mechanical processes present no problem. The alpha-beta method can be applied. The evolutionary process presents some problems because it implies the existence of a threshold value at which the endogenous variables break with the past. This prediction can make the model unfalsifiable: if the threshold value has been observed, the model is accepted; if it has not been, it is also accepted, with the argument that it is not time yet for the threshold value to appear and will be observed in the future. But the future is unobservable. An example is found in physics: The laws of gravity predict a big crunch (the collapse of whole universe) in the future.

What can be falsifiable is the dynamic segment of the evolutionary model, which is provided by the trajectory moving towards the threshold value that is predicted by the model. This is certainly a beta proposition. If this trajectory is refuted by facts, then the evolutionary model will be rejected, otherwise it will be accepted provisionally, as the rule of theory acceptance indicates for all cases.

In sum, the equilibrium conditions of a theoretical model may or may not be observable. If it is observable, then the model derives empirical predictions. If we observe in the real world situations that do contradict the equilibrium conditions, the model will be refuted. Thus from a theoretical model two types of *empirical predictions* can be derived:

- (a) Observable equilibrium conditions of the endogenous variables
- (b) Causality relations between endogenous and exogenous variables

Both cases will be defined as beta propositions. They are derived from the assumption of the existence of equilibrium, and stable equilibrium, in the economic process. They make the model falsifiable or refutable.

The assumption of the existence of equilibrium is not a protective belt of the theory; on the contrary, it increases its chance of refutability. If statistical analysis shows consistency between beta propositions—including type (a) and type (b)—and empirical data, the abstract society constructed by the theory is a good approximation of the real world society; otherwise the model fails. If all models fail, then the theory fails, as shown above.

The study of capitalism that we are going to carry out in this book will utilize the alpha-beta method, which is a particular method of Popperian epistemology, as shown above. Therefore, we have established the rules of scientific knowledge that we are going to apply for rejecting or accepting economic theories that seek to explain capitalism. These rules are summarized in Figure 1.3. They will be seen in action in the rest of the book.

Figure 1.1. Diagrammatic Representation of a Process Analysis

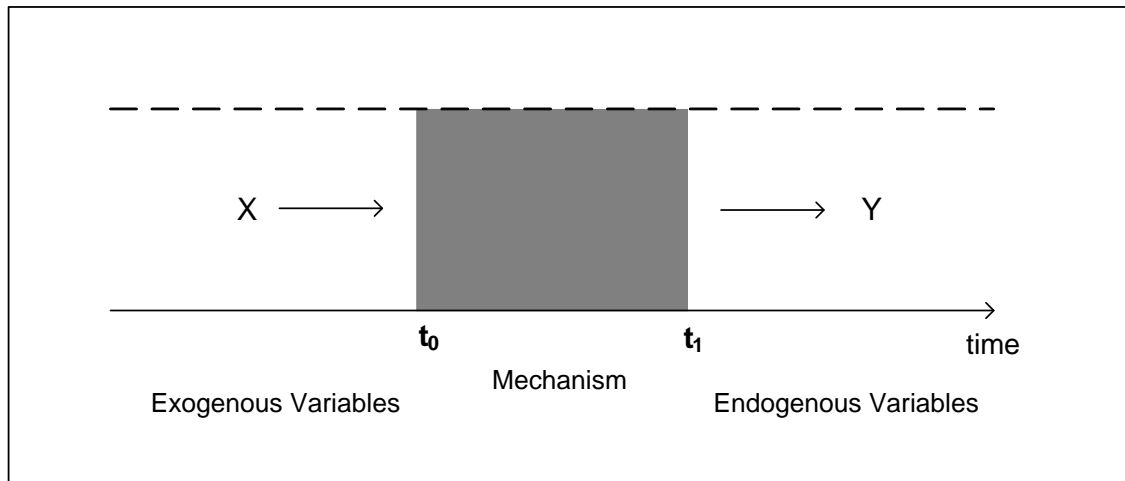


Figure 1.2. The Alpha-Beta Method
$$\alpha_1 \Rightarrow \beta_1 \rightarrow [\beta_1 \approx b]$$

If $\beta_1 = b$, α_1 is consistent with b and explains reality

If $\beta_1 \neq b$, α_1 does not explain reality and is refuted by facts. Then,

$$\alpha_2 \Rightarrow \beta_2 \rightarrow [\beta_2 \approx b]$$

If ... (continue with the same algorithm)

Figure 1.3. The Alpha-Beta Method in Economics

$$\alpha - (A') \rightarrow \alpha' \Rightarrow \beta' \rightarrow [\beta' \approx b]$$

If $\beta' = b$, then α' explains and so does α

If $\beta' \neq b$, then α' fails to explain; then

$$\alpha - (A'') \rightarrow \alpha'' \Rightarrow \beta'' \rightarrow [\beta'' \approx b]$$

If ... (the algorithm is followed)

$$\alpha - (A^n) \rightarrow \alpha^n \Rightarrow \beta^n \rightarrow [\beta^n \approx b]$$

If $\beta^n \neq b$, then α^n fails to explain, so does α . Then, a new theory is constructed and the algorithm continues.

CHAPTER 2

PRODUCTION AND DISTRIBUTION UNDER CAPITALISM: EMPIRICAL REGULARITIES

This book seeks to present the production and distribution economic laws in capitalist societies and then to explain them. What is capitalist society? What are those economic laws? This chapter seeks to answer empirically both questions.

2.1 Scope of the Book: Capitalist Countries, 1950-2010

The empirical definition of capitalist societies will include those independent countries that have been operating for long periods under the institutional norms of capitalism, which include: (a) private property of physical capital is more predominant than state property; (b) the market system is more predominant than non-market forms of exchange of goods between individuals; (c) democracy is the predominant form of governance, although it includes different degrees of people participation in public policies.

The period of analysis will be limited mostly to the period 1950-2010. This selection is due to the constraint given by the availability of statistical information.

According to our definition, and using the World Bank statistics, there are 174 capitalist countries in the world (World Bank 2001, pp. 334-335). They will be divided into two groups: the First World and the Third World. The First World consists of the richest 23 countries, according to the World Bank estimates (Ibid, Table 1, pp. 274-275). They are: Australia, Canada, Japan, New Zealand, USA, and the 18 countries of Western Europe (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and United Kingdom). The rest of 151 countries constitute the Third World.

Just for the sake of completeness, we should notice that socialist or communist countries are in turn characterized by the predominance of state property of physical capital over private property, the predominance of state planning over markets, and authoritarian political regimes in the period 1950-2010. This category, called the Second World, consists of 33 countries. They can also be separated into two groups. The first group consists of countries “in transition” to capitalism due to the introduction of market reforms since the 1990s, which include the countries of Eastern Europe, the Balkans, and the ex-Soviet Union. In the second group, we find five countries that have remained as communists in the period of analysis, and include China, Cuba, Laos, North Korea, and Vietnam.

The scope of this book is the study of the capitalist system only: production and distribution in the First World and the Third World, as defined above. The book is not about production and distribution in the entire world economy. Therefore, it should be clear from the outset that the Second World countries, as defined above, are not part of the analysis of the book. In particular, the reader should keep in mind that China is not part of the scientific analysis presented in the book.

Given its large population size and the rapid economic growth in the last two decades, China has changed the level and the structure of the world economy. But China is not a capitalist country even today and cannot be introduced into a theoretical system that assumes the capitalist form of production and distribution. China's economic growth has increased the size of world output and international trade, but it has not changed the rules under which capitalist countries operate. So China can safely be ignored in the analysis of capitalism over the past six decades. However, in the analysis of capitalism, the phenomenon of China's rapid growth will be taken into account, but as an exogenous factor. The empirical data on production and distribution in the capitalist countries presented in the book will include the China effect, although this effect will not be singled out.

2.2 A Brief History of Capitalism and Colonialism

The origins of capitalism are usually traced to the industrial revolution that took place at the end of the 18th century and beginning of the 19th century in Western Europe. If we take 1800 as the initial point, then capitalism has operated for nearly two centuries now.

How have capitalist countries come about? The well-known proposition attributed to Karl Marx says that societies tend to march along the same path in their history. "Marx saw all countries as following the same path from slavery to feudalism to capitalism, and finally to socialism and communism" (Nolan 2006, p. 353).

Consider a simplified trajectory in which societies start with a primitive or natural form of production and then move to more complex forms, beginning with the feudal form of production, characterized by agrarian societies, in which landlords concentrate the property of land and thus use landless workers for production under bondage systems; then comes the capitalist form of production, characterized by industrial societies, in which capitalists concentrate the property of physical capital and hire free workers in the labor market; the final stage will take the form of a communist society. To these types of human societies, European colonialism is introduced.

Analytically, therefore, the evolution that today's world countries have experienced could be summarized in a stylized manner using those categories, since the discovery of America. This is shown in Table 2.1. Current countries of the world are grouped into nine categories. The first five groups correspond to capitalist countries; groups (6) refers to what is called countries "in transition" from communism to capitalism, whereas the last three refer to communist countries.

Some comments on Table 2.1 are in order. Colonialism refers to the domination of societies by Europeans in other regions of the world. The domination of a society upon others within the same region, either within Europe or within Asia, is not part of

this definition. The idea is that the domination of the New World societies by Europeans, which were very different in technology and culture, is what becomes significant in the postcolonial period, as historical legacy.

A distinction is made between settlement and colonialism. The category utilized for the cases (2) and (3) is settlement, in which the European domination operated on empty lands or lands with very low population density. In the New World, USA and Canada were not colonies, but rather settler territories, a distinction made by historians. "North America was a different matter. Here there were no Amerindian settlements of sufficient size to be worth conquering and consequently no firm bases for further exploration. The conquistador De Soto found few Indians and no gold" (McEvedy, 1972, p. 20). The same statement could be applied to Australia and New Zealand and also to Argentina, Costa Rica, and Uruguay in Latin America.

Group (1) includes Western Europe and Japan. Some of the European countries were colonial powers (England, Spain, Portugal, France, Germany, Holland, and Belgium). Group (2) includes countries that were settlements, not colonies, and belong to the First World.

Group (3) shows three countries (Argentina, Costa Rica, and Uruguay) that were also settlements but do not belong to the First World, but to the Third World. This is out of the pattern shown in case (2). Indeed, historian Angus Maddison (1995) pointed out that in the history of capitalism there are two special countries. Japan is the only country that has moved into the First World group; Argentina is the only country that has left the First World group. Group (4) includes those countries of the Third World that were never European colonies, only nine countries. In contrast, group (5) shows that most Third World countries have a history of European colonialism.

This classification intends to be general, ignoring some details. For example, some countries of the Third World were under communist regime for short periods. However, considering countries that were under a communist regime for 10 to 17 years, the longest periods recorded, the list includes only six countries: Angola, Benin, Congo, and Mozambique in Africa, and Afghanistan and Cambodia in Asia.

As can be seen in Table 2.1, the predictions of Marxian theory are refuted by the historical facts: the sequence *Naturalism-feudalism-capitalism-communism* has not been observed in any country of the world. It seems that Marx assumed the reality of Western Europe, which is *Naturalism-feudalism-capitalism*, as the typical evolution and extrapolated to the rest. Colonialism has certainly played a significant role in shaping the observed sequence. In most Third World countries, case (5), capitalism has emerged from colonialism.

In sum, from the 151 countries classified as Third World above, only nine were not European colonies. Afghanistan, Iran, Ethiopia, Thailand, and Turkey have never been subject to colonial domination; other three countries have not been colonized by the Europeans, but they have by the Chinese and Japanese, although for a short period of time only (less than 50 years): Taiwan, South Korea, and Singapore (Wesseling, 2004). If we add to this group of eight countries that have never been colonized by Europeans, the three countries of Latin America that were mostly settlements, group (3), and include the more recent country of Israel (founded in 1948), then twelve Third

World countries have none or weak colonial legacy. Therefore, the large majority of Third World countries, 139 out of 151, have strong colonial legacy.

The scope of this book is the study of capitalism. Groups (1) through (5) in Table 2.1 are then the relevant categories for our study, 174 countries in total. Groups (1) and (2) refer to what we have defined here as the group of First World countries, 23 in total. The 151 countries of the Third World comprise groups (3) and (4), which constitute a group of 12 countries that will be called *Third World with Weak Colonial Legacy*, whereas group (5) constitutes the *Third World with Strong Colonial Legacy*, composed of 139 countries. These two groups of Third World countries will become analytically important in the rest of the book.

2.3 Empirical Regularities on Production and Distribution

Once the types of capitalist countries have been defined, the aggregate behavior of these three groups of capitalist countries is now presented as a set of empirical regularities. These regularities refer to production and distribution variables. They are statistical regularities, that is, they are valid in the vast majority of countries, or in the average, but not necessarily in all countries. The remainder of the book will present theories that intend to explain these regularities.

Estimates of the differences in average income and average degree of income inequality in these three groups of capitalist countries are now presented in Table 2.2. It is shown there that the average income of the First World is much higher than that of the Third World with strong colonial legacy, whereas the Third World with weak colonial legacy lies in between. This result is just confirming our definitions: the richest countries are called First World and the poorest countries Third World, as an objective way of describing this ranking.

The degree of income inequality between individuals will be measured by the Gini index, which varies between zero and one. The value of zero indicates perfect equality in the distribution of income, whereas the value of one indicates perfect inequality (all income concentrated in the hands of one individual). Certainly, the empirical Gini index can hardly reach the value of zero or one. The observed empirical values calculated in the international literature hardly reach beyond 0.70; on the other hand, the lowest values observed rarely go below 0.20. Therefore, the use of the terms “low-inequality” and “high-inequality” societies in this book will be in reference to these actual extreme values. For each country, the total income that is distributed is usually total household income of the country.

Table 2.2 shows that the within-country inequality is, on average, lower in the First World compared to what it is in the Third World with strong colonial legacy, while the value in the Third World with weak colonial legacy lies in between. It is clear that the First World is not only wealthier than the Third World, but it also more equalitarian. These relative differences in levels constitute important traits of the capitalist system.

The production and distribution regularities within and between the First World and the Third World in the period 1950-2010 for which statistical data are available can be summarized as follows:

Fact 1: The Existence and Persistence of Unemployment in the First World

First World countries generally operate with unemployment. This is an undesirable situation for workers. Unemployment rates may vary by countries and by periods within a country, but these rates are always positive. The U.S. economy showed average annual rates of unemployment that vary from 3% to 10% in the period 1960 to 2005, whereas in Western Europe these rates were 2% and 11% in the same time period (Blanchard, 2009, Table 1.5). The new feature of the ex-socialist economies in transition to capitalism is the fact that unemployment has appeared; that is, capitalism and unemployment arrived together. Unemployment may thus be considered the disease of capitalism, as stated by the historian John Garraty (1978).

Fact 2: The Existence and Persistence of Unemployment and Underemployment in the Third World

Three groups of workers can be distinguished in the Third World. The first is employed in firms, who are paid market wages. The second is the unemployed. Unemployment rates are generally positive in Third World countries. These rates also show variations by countries and by periods, and with rates in the range that is similar to that of the First World. The third is the group of workers that are self-employed in small businesses and small farms (with no hired labor) where they generate their incomes.

Self-employment rates are more significant in the Third World than in the First World. There are estimates for urban areas that show this difference. According to ILO statistics, in the period 1980-2000, the rate of self-employment in the First World was 12% on average (over a sample of 17 countries), whereas in the Third World it was 40% (over a sample of 58 countries) (ILO 2002, Annex 2, p. 62-64). If the rural areas were included, this gap would be even higher, because self-employment rates are much higher in rural than urban areas and this gap is probably higher in the Third World.

Empirical studies show that the average income of the self-employed in the Third World is smaller than the average wage prevailing in the labor market, for a given level of skills and a given working-length period—fulltime(cf. Figueroa 2010, Telles 1993). The situation of the self-employed is therefore involuntary and undesirable, as in the case of the unemployed, because these workers would prefer to get wage employment. This group will be called the *underemployed*.

The excess labor supply takes the form of unemployment in the First World, whereas in the Third World it is composed of unemployment and underemployment, the latter being the more significant. Therefore, the common practice of using the unemployment rate as the criterion for making international comparisons about the excess labor supply is unwarranted. Unemployment

figures underestimate the magnitude of the total excess labor supply in the Third World.

Fact 3: The Existence and Persistence of Income Gaps between Ethnic Groups in the Third World

Most Third World countries are multiethnic. Empirical studies on the income gaps between ethnic groups in countries of Latin America, Africa, and Asia have found the existence and persistence of these gaps (Darity & Nembhard, 2000; Figueroa, 2010; Hall & Patrinos, 2005; Psacharopoulos & Patrinos, 1994; Silva, 2001; Stewart, 2001; & Telles, 1993). In explaining the overall inequality in the Third World countries, ethnicity of social groups does seem to matter.

These empirical studies also show that gaps are systematic, that is, ethnic groups maintain their positions in the income pyramid of their respective countries. In particular, those ethnic groups that are descendants of aboriginal populations or slave populations, the legacy of colonial systems, have low average incomes compared to the national average. Indigenous peasants, for instance, constitute almost invariably the poorest social group in every country of the Third World.

Fact 4: In the short run, nominal and real variables are correlated in both the First World and the Third World

In the workings of capitalist countries, empirical studies on the relationships between changes in the nominal variables under the control of governments (money supply, exchange rate, interest rate) and real variables (total output, real wages and employment) in the short run have found that they are correlated. For instance, monetary aggregates tend to be procyclical (they move in the same direction of total output) in the U.S. economy, based on quarterly data from 1959 to 1996 (Barro, 1997, Table 18.1, p. 705). Money aggregates have also been found to be procyclical in a sample of 10 countries of the First World and of 15 countries of the Third World (from Sub-Sahara Africa, Latin America, Asia, and North Africa) and for the period 1970-1997 (Rand & Tarp, 2002).

Fact 5: In the long run, real wage rates and output per worker are positively correlated in both the First World and the Third World

When output per worker grows for long periods, real wages also tend to increase. For example, in the U.S. economy, in the period 1950-1974, output per worker grew at the annual rate of 1.9% per year, while real wage grew at a similar rate (Barro, 1997, chap. 6, pp. 225-226). In the Third World, a sample of 12 Latin American countries showed positive growth rates in both output per

capita and real wages in all but three countries in the period 1980-1999 (ILO, 2000, Table 9A).

Fact 6: In the long run, the income level gap between the First World and the Third World is persistent over time

As already depicted in Table 2.2, the average income level of First World countries is higher than in the Third World in 2008. This does no more than checking our definition. What is significant is that this gap has shown persistence over the period of study. Empirical studies have measured the persistence of this income gap by estimating growth rates of output per capita across countries. If these rates were higher for the Third World, one would expect a catching up at some period in the future; if the relation were on the contrary, there would be no chance of catching up. The empirical literature has shown that the latter case is the common finding. Thus the classical study by Barro and Sala-i-Martin (2004) found that per capita output in poor countries did not grow faster than in rich countries in the period 1960-2000. Persistence of income inequality between First and Third World countries is thus another empirical regularity of capitalism.

For the very long run, the estimates made by economic historian Angus Maddison indicate a drastic *increase* in the income gap. The income per capita of the richest countries compared to the poorest was 3:1 in 1820, which jumped to 18:1 in 2000 (cited in Galor 2011, Figure 1.1, p. 2).

Fact 7: In the long run, the degree of within-country inequality in the Third World is, on average, higher than in the First World, and this difference is persistent over time

The group of First World countries is, on average, more equal in the within-country distribution of income compared to the group of the Third World countries, as shown in Table 2.2. Over time, these differences tend to persist. In fact, empirical studies have shown that the Gini index for both the First World and Third World countries shows significant viscosity in the period 1950-1995 (e.g., Atkinson 1996, Deininger & Squire, 1996; Li, Squire, & Zou, 1998).

Fact 8: In the very long run, the process of economic growth is accompanied by the degradation of the biophysical environment

The economic growth experience of the capitalist system in the period 1820-2001 can be approximated by the calculations made by economic historian Angus Maddison for the world as a whole. Total output grew at the average rate of nearly 1% per year in the period 1820-1950, but at the rate of 4% in the period 1950-2001 (Maddison 2003, pp. 256-257).

The global economic growth experience in the last two centuries has been accompanied by an equally rapid degradation of the environment. This period has been accompanied by an equally rapid degradation of the environment, which includes depletion of non-renewable resources and pollution. Some studies show that some non-renewable resources are getting near exhaustion (Clugston 2012). The standard measure of pollution is the concentration of greenhouse gases in the atmosphere, such as carbon dioxide (CO₂), methane (CH₄), and nitrous oxide (N₂O). The measurement of these concentrations over the past 2,000 years confirms the large increases during the last 200 years, and most of which occurred during the 20th century, and are attributed to anthropogenic emissions (Etheridge et al 1996 and MacFarling 2006).

It is clear that the time path of pollution of the environment coincides with the time path of total output growth. Physicist Richard Muller makes the connection as follows: “The amount of carbon dioxide was pretty constant from 800 AD until the late 1800s, at the level of 280 ppm. In the last century it has shot up to 380 ppm—an increase of 36%. If we continue to burn fossil fuels, we expect the carbon dioxide to keep rising....The carbon dioxide comes from human activity, including burning of fossil fuels and the destruction of enormous regions of forest” (Muller, 2008, p. 265-266). (CO₂ concentrations are measured as parts per million, abbreviated ppm.)

2.4 The Need of Theory to Explain the Empirical Regularities

The scientific question now is to see whether there exists an economic theory that can explain the eight regularities altogether. If it does, such theory will explain why capitalism functions in this manner.

Neoclassical theory and Keynesian theory constitute the standard economics of today. These two theories complement each other. Keynesian theory is utilized in the short run analysis (inflation, unemployment, recession) and the neoclassical theory in the long run analysis of economic growth. The analytical reason for the complementarity is that Keynesian theory becomes neoclassical theory in the long run analysis. Although there is a debate between these two theories over their capacity to explain the short run, the concept of standard economics for the long run analysis, with which this book is mostly concerned, is still valid. As will be shown in Chapter 3, standard economics is not able to explain the full set of regularities established above. Classical economic theory also fails to explain these regularities. Consequently, a new theory is needed and one will be presented in this book.

The new theory will assume the existence of three different types of capitalist societies, which are based on their initial conditions or history. It will firstly seek to explain each type of capitalism as defined above, but taken separately; then it will seek to explain the capitalist system taken as a whole. As the book will show, this theory, called *unified theory*, will be able to explain the eight empirical regularities.

Old and new economic theories to explain capitalism are then presented in this book. The rules of scientific knowledge that will be used to reject and accept theories

throughout the book correspond to those of the Popperian epistemology, which was developed in the previous chapter. Standard economics and classical economics have a significant heuristic value in the construction of new theories. Surely, one can make progress more easily if one is standing on the shoulders of giants, and giants were those who constructed and developed the science of economics up to now. Therefore, the new theory presented in this book will not be fully understood unless the foundations of the received economics are understood. Those foundations are presented in the next chapter.

Table 2.1. Evolution of Countries to Capitalism, 1500-2010

- (1) Western Europe and Japan: *Naturalism—Feudalism—Capitalism*
- (2) USA, Canada, Australia, and New Zealand: *Settlement—Capitalism*
- (3) Argentina, Costa Rica, Uruguay: *Settlement—Capitalism*
- (4) Never were colonized by Europeans (Afghanistan, Ethiopia, Iran, Israel, Singapore, South Korea, Taiwan, Thailand, Turkey): *Naturalism—feudalism—Capitalism*
- (5) Most of Africa, Asia, and Latin America: *Naturalism—Colonialism—Capitalism*
- (6) Eastern Europe (Ex-Soviet Union, Eastern Europe, Balkans, Mongolia): *Naturalism—Feudalism—Communism—Capitalism*
- (7) China: *Naturalism—Feudalism—Communism*
- (8) Cuba, Laos, Vietnam: *Naturalism—Colonialism—Capitalism—Communism*
- (9) North Korea: *Naturalism—Colonialism—Communism*

Source: Elaborated by the author from Dalziel (2006) and Wesseling (2004).

Table 2.2. Income Level and Income Inequality in Capitalist Countries

Group	Average Income		Income Inequality	
	US\$, PPP		Average Gini Index	
	1980	2008	1950-1970	1971-2008
A. First World	9,508 (1,977) 23	38,563 (9,570) 23	0.36 (0.04) 15	0.33 (0.04) 22
B. Third World with Weak Colonial Legacy	3,779 (2,052) 9	19,473 (13,312) 9	0.39 (0.09) 5	0.36 (0.07) 7
C. Third World with Strong Colonial Legacy	2,360 (3,662) 65	6,022 (6,544) 72	0.47 (0.09) 30	0.49 (0.07) 32

Notes:

Figures in parenthesis indicate standard deviation and those in the third row number of countries in the sample. Total number of countries in each category is 23, 12 and 139.

Average income: It is Gross National Income per capita, adjusted by purchasing power parity (PPP) to reflect differences in the purchasing power of the American dollar in different countries relative to US. Samples of First World Countries include 23: Australia, Canada, Japan, New Zealand, USA and the 18 countries from Western Europe; of Third World Countries with Weak Colonial Legacy include 9: Argentina, Costa Rica, Iran, Israel, Singapore, South Korea, Thailand, Turkey, Uruguay; and of Third World Countries with Strong Colonial Legacy include a total of 73.

Gini Index: Considers only comparable data across countries: the distribution of net income per capita calculated from national household surveys. Third World Countries with Weak Colonial Legacy include Costa Rica, Israel, Singapore, South Korea, Taiwan, Thailand, and Turkey.

Sources:

Groups of countries: World Bank (2001, pp 334-335); group B from Dalziel (2006) and Wesseling (2004). Average income: Calculated from World Bank - Data. <http://data.worldbank.org/indicator/NY.GNP.PCAP.PP.CD/countries>. Last date of revision: Not indicated. Last date accessed: 19/10/2011. Gini Index: Calculated from the data base of Branko Milanovic (2010), Web Page at World Bank, "All the Ginis Dataset": <http://go.worldbank.org/9VCQW66LA0>. Last date accessed: 19/10/2011.

CHAPTER 3

STANDARD AND CLASSICAL ECONOMICS

Several theories that seek to explain production and distribution in the capitalist society coexist in economics. Neoclassical, Keynesian, and classical theories constitute the most important ones. Are these theories able to explain the eight empirical regularities of capitalism? The foundations of these theories are firstly presented in this chapter; then models are developed and empirical predictions are derived, which are confronted against the empirical regularities.

3.1 The Neoclassical Society

The abstract neoclassical society or neoclassical theory is presented in this section. This construction is inspired in the school initiated by Adam Smith (1776) and Leon Walras (1883).

The primary assumptions of neoclassical theory are the following. As regards to the institutional context, in this society people participating in the economic process are endowed with quantities of economic assets. Assets are goods that provide a flow of incomes.

People exchange goods in the form of *market exchange*. Market exchange is a particular mechanism for exchange of goods among individuals and is subject to some rules, which include:

- (a) It is voluntary.
- (b) It is based on the motivation of self-interest, which implies that people seek to take advantage of every exchange possibility to make economic gains for their own benefit.
- (c) It is constrained by the individual's endowments of economic assets only, which means freedom of exchange, absence of social or legal restrictions.

An implication of market exchange is that it is impersonal. The individual seeks personal gains from exchange. Therefore, the law of one price for buyers and sellers will prevail in each market.

By contrast, non-market exchange of goods may take the form of *reciprocity*. It refers to exchange of goods and favors, in which social assets and social norms play a role. In this case, pure market exchange rules (a) and (b) do prevail; but (c) does not. The implication is that personal relations are essential for the exchange of goods. The

exchange balance need not hold in the short run; it will be attained in the long run; hence, the law of one price need not apply in the short run. This form of exchange is restricted to social networks, such as families, social groups, and communities.

In the neoclassical society, people exchange goods through the norms of market exchange. (Non-market exchange is not essential and can thus be ignored.) All markets are Walrasian. A *Walrasian market* is a form of market exchange in which individuals are able to buy or sell the good under exchange in the quantities they desire at the prevailing market price; that is, no one is left without realizing the desire transaction. If this situation were not attained, the price of the good would change, until the market clears. A Walrasian market assumes fully flexible prices. Price plays the role of a market rationing mechanism: if the quantity demanded of a good is higher than the quantity supplied, then the market price will increase; if it is the other way round, then the market price will fall. The assumption is that the market system is self-regulated; that is, it operates *as if* it solved a system of equations.

The alpha propositions of the neoclassical theory (N) are the following:

$\alpha(N)$. (1) *Institutional Context*: (a) Rules: People participating in the economic process are endowed with economic assets, which are subject to private property rights; people exchange goods subject to the norms of market exchange. The market system is composed of Walrasian markets. The political regimen is democratic. (b) Organizations include households, firms, and the government.

$\alpha(N)$. (2) *Initial Conditions*: Factor endowments of society (capital per worker) are such that the productivity of total labor is high enough to cover wages and generate profits.

$\alpha(N)$. (3) *Economic Rationality of Agents*: Individuals act guided by the motivation of self-interest.

A Neoclassical Model: The Short Run

In order to make neoclassical theory falsifiable, a neoclassical model must be constructed, which can be seen as setting a particular stage in which social actors will play their roles. The following auxiliary assumptions are introduced for that purpose:

- There are two social groups: capitalists who are endowed with stocks of physical capital and human capital, and workers with stocks of human capital and cash balances.
- Government behavior consists of supplying money to the economic process.
- Three markets—all Walrasian—will constitute the market system: labor, commodity, and money. Workers are endowed with the same human capital. Hence there is only one labor market. One good is produced in society, called good B. Workers and capitalist firms exchange labor services and goods in the labor and commodity markets; money supplied by the government is used as the means of payment and is exogenously

determined. Wages are paid at the end of the production period; consequently, workers need to hold cash balances for transaction purposes. There is no market for renting capital services.

- Capitalists and workers interact and compete in the market to obtain their objectives of self-interest, but none has the power to set prices (perfect competition); they all face costless information on market prices and technology. Cost curves are different among firms.
- The economic process is static; it is closed as well, for market exchange with other societies is ignored. Short run equilibrium is analyzed.

“Short run” is a logical category, as it implies that the productive capacity of society is given. However, it has an implication for the period of analysis as well, as it must correspond to an economic process that is repeated in short periods (quarterly or yearly), in which variations in the productive capacity will be small and thus may safely be ignored. Endogenous changes in the productive capacity will be analyzed through “long run” equilibrium models.

With these assumptions the neoclassical theory becomes operational. Beta propositions can be derived from this theoretical model, which can be confronted against the empirical regularities of capitalism.

The Behavior of Workers

In the analysis of the economic behavior of social actors, consider firstly the social group of workers. Workers seek to maximize the quantity of good B for consumption, subject to their budget constraint and their real cash balance constraint. Workers need to hold a quantity of real cash balance for transaction and precautionary purposes. The stock of money to be held is a requirement to make transactions in the market, according to the practices of payments. The real cash balance depends upon the real income: the higher the workers’ income, the higher the real cash balance required. But this requirement is not proportional, the model assumes that double income requires less than double real cash balance.

Therefore, for given real income, there is an optimum stock of money that workers must hold. If the cash balance held is higher than the required, workers will seek to run down the stock by spending the excess in the purchase of good B; if the cash balance is smaller, they will build up the stock by buying a smaller quantity of the commodity. The adjustment mechanism to get the optimum cash balance operates via changes in the quantities of good B bought. Assume this adjustment in stocks is made in one period. Once this adjustment is made, workers will buy a given quantity of good B that their wages permit, the equilibrium quantity, period after period.

Workers exchange their labor for good B in the market. They are able to sell all their labor in the labor market at the prevailing wage rate and are also able to buy all the quantity of good B that they are willing to exchange because they exchange these goods in Walrasian markets.

Workers are price-takers in these markets; that is, they take nominal prices (P_h and P_b) as given. These are the exogenous variables, upon which workers cannot decide. Their endowments are also exogenously given: labor (S_h) and nominal cash balance (S_m). The quantity demanded of good B (D_b) and the real cash balance (D_m/P_b) are the endogenous variables, upon which workers can decide.

The behavior of the individual worker i can be represented by the following system of structural equations:

$$\begin{aligned} P_h S_{hi} &= P_b D_{bi} + (D_{mi} - S_{mi}) \\ D_{mi}/P_b &= L_i(D_{bi}), L_i > 0, L_i' < 0 \end{aligned} \quad (3.1)$$

The first equation shows the worker's budget constraint. The second equation is the demand for real cash balances. The equilibrium condition is that the individual's cash balances are willingly held, that is, $(S_{mi} - D_{mi}) = 0$. As long as the exogenous variables remain fixed, workers will maintain the values of the endogenous variables constant period after period. This situation represents a static equilibrium.

This equilibrium is (trivially) stable. Therefore, the method of comparative statics can be applied to derive beta propositions, that is, to determine the effect of exogenous variables upon endogenous variables.

An increase of the nominal price of good B would have the following effects. Real wages would fall and the quantity demanded of good B would also fall. At this lower level of real income, the required level of real cash balances would fall. But the rise in the price would lead precisely to a fall in the initial real cash balances. Then the new and old stock would tend to equalize. Of course, if the relation between real income and real cash balances were proportional, these values would indeed equalize; but the relation is not proportional, as workers would end up with a deficit of real cash balances. They would have to increase the stock of money, which implies buying a smaller quantity of good B in the period of adjustment. There would be a *real-cash balance effect* on the demand for good B. Once this adjustment is made, the new equilibrium implies lower real income and consumption of good B, together with a lower amount of real cash balance, period after period.

Consider now a rise in the nominal wage. The quantity of good B demanded would rise, while the real cash balance would remain unchanged. The required real cash balances would increase and workers would experience a shortage of real cash balances. The adjustment of this stock would imply buying, in the period of adjustment, a smaller quantity of good B than the new equilibrium value. Once this adjustment is made, the new equilibrium implies that the worker would have a higher real income, consume a higher amount of good B, holding a higher amount of real cash balance, period after period.

Lastly, consider an increase in the worker's money endowment. This increase generates excess in his holdings of cash balances. The worker will seek to get rid of this additional quantity of money by buying an extra quantity of good B. This adjustment is assumed to be made in just one period; hence, in the adjustment period, the worker will buy an additional amount of good B. Once this adjustment is made, the equilibrium quantity of cash balances is restored and the worker will continue to consume the initial

amount of good B, holding the initial nominal (and real) cash balance, period after period.

The real cash balance effect of a change in the price level or in the nominal wage rate upon the quantity of good B demanded operates only in the adjustment period. Over several periods, where the equilibrium values prevail, this effect will be small and may be ignored. A change in the money endowment of the individual worker has a similar effect as the real cash balance effect. The effect upon the quantity of good B demanded is positive but appears in the adjustment period only; that is, over several periods the effect is small and may be ignored. Hence, real-cash balance effects can be safely ignored.

On the behavior of the individual worker, in sum, changes in the values of the exogenous variables change the values of the endogenous variables to a new equilibrium situation. The result is that the quantity demanded of good B coming from the individual worker depends directly upon the nominal wage rate and inversely upon the price level. If both nominal prices increase in the same proportion, which means maintaining the real wage rate fixed, the quantity demanded will remain unchanged; hence, the quantity demanded of good B will depend upon the real wage rate only. This relationship can be represented as the individual worker's demand function for good B as follows:

$$D_{bi} = f^i(P_h, P_b, S_{hi}) = F^i(w, S_{hi}), \quad F^i_1 > 0, F^i_2 > 0 \quad (3.2)$$

The Behavior of Capitalists

Consider now the behavior of the social group of capitalists. Capitalists are endowed with stocks of physical capital (and the homogeneous human capital as well). Capitalists produce good B with machines and workers in the firms they own. Technology is given and known to all. Two other factors that affect output are also considered as given: length of the working period (say, 40 hours per week) and degree of effort intensity of workers.

Therefore, capitalists know that the quantity of output per unit of time will depend upon the quantity of workers hired; they also know that more workers will produce more output, but double number of workers will not produce double quantity of output, but less. This is due to the assumption that production is subject to *diminishing returns* of labor inputs. Output per worker is also called *average productivity of labor*. Diminishing returns implies that average productivity of labor declines as more workers are hired. The additional output that results from additional worker is called *marginal productivity of labor*. Diminishing returns also implies that the marginal productivity of labor declines as more workers are hired.

Firms seek to maximize total profits. Profits are equal to the difference between total revenues and total costs. Hence profit maximization implies maximizing total revenues and minimizing total costs. Total revenues are equal to the price of good B multiplied by the quantities of good B sold; total costs are equal to the wage bill (nominal wage rate multiplied by the quantity of workers hired) plus the cost of depreciation of physical capital.

The cost of depreciation is a fixed cost, for it is a fixed proportion of the value of the capital stock of the firm. The cost of depreciation allows the firm to maintain the stock of capital constant period after period. Depreciation is what makes production a process. Therefore, profits will be net of depreciation in what follows and total output will also be net of depreciation.

Consider now the behavior of capitalists taken individually. Individual capitalists are price takers in both markets: good B and labor. Assume firms have cost differences due to differences in the entrepreneurial talents of capitalists and to differences in the firm's location and access to public infrastructure. Because they operate in Walrasian markets, they can sell all the quantities of good B and can buy all the quantities of labor that they are willing to exchange at the given market prices. Capitalists do not hold cash balances. There is no need for working capital, for they pay wages at the end of the production period.

Each firm will seek to hire the number of workers that makes profits the largest; that is, the number that makes the difference between revenues and costs the largest. The behavior of the individual capitalist j can be represented by the following structural equations:

$$\begin{aligned} &\text{Maximize} && P_j' = P_b Q_{bj} - P_h D_{hj} && (3.3) \\ &\text{Subject to the production function} \\ &&& Q_{bj} = \Phi_j(D_{hj}, K_{bj}), \Phi_{j1} > 0, \Phi_{j11} < 0 \end{aligned}$$

This is a mathematical problem that the capitalist will seek to solve by different methods, such as intuition, trial and error, and so on. The assumption is that, whatever the method used, the solution found by the capitalist will be equal to the mathematical solution. So, the capitalist behaves *as if* he solved this mathematical problem.

The mathematical solution indicates the equilibrium condition, which is the following:

$$\begin{aligned} &\Phi_{j1}(D_{hj}, K_{bj}) P_b = P_h \quad \text{or} \\ &\Phi_{j1}(D_{hj}, K_{bj}) = w \end{aligned} \tag{3.4}$$

The first equation indicates that profit is maximized when the value of the physical marginal productivity of labor is equal to the nominal wage rate. The second is obtained just by dividing the first by P_b , and indicates that profit is also maximized when the physical marginal productivity of labor is equal to the real wage rate. Either equation has only one variable to solve for: the quantity of labor demanded by the firm (D_h), which is the number of workers that will make profits the largest.

According to the theory, labor productivity is an important factor in the decision of firms to hire workers. It should be emphasized, however, that productivity is not a personal characteristic of the worker, something that could be written in his ID. The same worker accompanied by 30 other workers will result in a certain quantity of total output, which will be different if accompanied by 60 workers. Labor productivity is a collective result. But labor productivity will also vary with different quantities of machines; or with different technology incorporated in the machines. In short, given the level of technology, labor productivity depends upon the number of workers and the stock of capital, as shown in equation (3.4).

The equilibrium condition is stable. If for some reason the firm has hired more workers than the optimal number, the nominal wage rate will be higher than the value of the physical marginal productivity of labor. Profits are not being maximized; then firms will reduce workers to increase the value of the physical marginal productivity (by the diminishing returns assumption) and the equilibrium number of workers will be restored. If for some reason the firm has hired fewer workers than the optimal number, by the same argument, firms will seek to hire more workers and the equilibrium number of workers will also be restored. Therefore, the comparative static method can be used to determine the effects of changes in the values of the exogenous variables upon the endogenous variable

If the price of good B goes up, then the value of the marginal productivity of labor will increase and it will become higher than the nominal wage rate; thus the quantity of labor hired by the firm will increase. If the nominal wage rate rises, the value of the marginal productivity of labor will become smaller than the nominal wage; then the quantity of workers hired by the firm will decrease. Finally, if the capital stock of the firm increases, then the value of the marginal productivity of labor will rise (because workers will be equipped with more machines) and will become higher than the nominal wage rate; hence, the quantity of labor hired will increase. The individual firm's labor demand function can then be written as follows:

$$\begin{aligned} D_{hj} &= h^j(P_h, P_b, K_{bj}) = H^j(w, K_{bj}), \quad H^j_1 < 0, H^j_2 > 0 \\ &\text{Subject to the firm's budget constraint} \\ P_b S_{bj} &= P_b Q_{bj} - P'_j = P_h D_{hj} \end{aligned} \quad (3.5)$$

Market General Equilibrium

Workers and capitalists interact through market exchange. The solution of such interactions is the object of general equilibrium analysis. There are three markets in this model: labor, good B, and money. Workers are willing to sell labor and buy good B; capitalists are willing to buy labor and sell good B. In addition, workers need to hold cash balances and the supply of money comes exogenously from the government.

The individual behavior of a typical worker was presented above. The aggregate demand function for good B and the aggregate demand function for money are obtained just by adding the individual demand functions. There are no direct interactions between workers, such as collective action; interactions among workers are only indirect, through markets. Hence, by aggregating equations (3.1) and (3.2) over all workers, we get

$$\begin{aligned} D_b &= F(w, S_h), \quad F_1 > 0, F_2 > 0 \\ D_m &= P_b L(D_b), \quad L > 0, L' < 0 \\ &\text{Subject to the aggregate budget constraint of all workers} \\ P_h S_h &= P_b D_b + (D_m - S_m) \end{aligned} \quad (3.6)$$

The individual behavior of firms was also presented above. There are no direct interactions between capitalists either, such as collective action; interactions among capitalists are only indirect, through markets. Hence, the aggregate demand function for labor is obtained just by adding the individual function, shown above as equation (3.5), over all firms, as follows:

$$D_h = H(w, K_b), H_1 < 0, H_2 > 0 \quad (3.7)$$

Subject to the aggregate budget constraint of all capitalists

$$P_b S_b = P_h D_h$$

In the market place, workers and capitalists are price-takers. Prices are exogenous to them. No one has the power to set the market prices. How are then market prices determined? Prices are determined by the interaction of all social actors. The equilibrium prices will be those that clear the markets: people willing to exchange quantities at those prices will be able to do so; if not, prices will increase or decrease, depending on excess demand or excess supply situations. Therefore, prices and quantities in the market system are endogenous and are determined by the interactions between buyers and sellers. This type of market interactions is called *competitive market*.

The general equilibrium conditions require that excess demand be equal to zero in each particular market. Then

$$\begin{array}{ll} \text{Labor market} & S_h = D_h \\ \text{Commodity market} & S_b = D_b \\ \text{Money market} & S_m = D_m \end{array} \quad (3.8)$$

Subject to the aggregate budget constraint

$$P_b (D_b - S_b) + P_h (D_h - S_h) + (D_m - S_m) = 0$$

The aggregate budget constraint, which is also a general equilibrium condition, is obtained just by adding up the budget constraints of the two social groups, workers and capitalists.

How does the competitive market system solve for prices and quantities? There are three markets and three nominal prices and three quantities to solve for, a total of six endogenous variables. There are two conditions in each market, a demand function and a supply function, a total of six conditions. The system thus contains equal number of equations and unknowns, so in principle a solution should exist.

In this particular neoclassical model, the unknowns are less than six. For one thing, the nominal price of money is one (the price of a dollar in dollars is one); two quantities are exogenously determined: quantity of employment and quantity of money. In addition, from the aggregate budget constraint, it follows that if two of the three markets are in equilibrium, the third will necessarily be in equilibrium too. This is known as the *Walras' Law of markets*, which assumes that markets are Walrasian.

Because the three markets are Walrasian, one of them can be eliminated from the general equilibrium analysis. We can therefore choose to eliminate the good B market as the redundant one; thus it is sufficient to solve the labor and money markets to reach general equilibrium. Because the quantities in these markets are exogenously determined; the core has two unknowns only (the price of good B or price level and the nominal wage rate) and two conditions (the demand function for labor and the demand function for money). Therefore, there are equal number of conditions to be satisfied and variables to be solved, that is, equal number of equations and unknowns in mathematical terms, which implies that in principle the solution should exist.

The assumption is that the market system is able to solve this problem. This is the general market theory. The market system works *as if* it solved a system of equations. In the real world the market system may use particular mechanisms to solve for prices and quantities (e.g., auctioneers), but these mechanisms are supposed to be equivalent to solving a system of equations.

The exogenous and endogenous variables of the market system are:

Exogenous variables: S_h , S_m , K

Endogenous variables: P_b , P_h , w , $D_b=S_b=Y$, P , W

In this particular model of the neoclassical theory, the general equilibrium solution is very simple. It can be solved graphically. Figure 3.1 depicts the labor market in panel (a) and the labor market in panel (b). In the labor market, due to the assumption of diminishing returns, physical marginal labor productivity decreases as more workers are employed to work with a fixed stock of capital (K), as shown by curve H , which also represents the labor demand curve. The labor market determines the real wage rate. Once the real wage rate is known, the real wage bill will be known, which will determine the level of the money demand curve, which in turn will determine the price level in the money market. Finally, once the price level is known, the nominal wage rate will also be determined so as to be consistent with the real wage rate.

Notice that the money market solution is determined once the labor market solution has been determined. The market system comprises two subsystems: the real one (the labor market) and the nominal one (the money market). In this particular model, the labor market constitutes the *core of the general equilibrium*: It is sufficient to solve this market because the rest of endogenous variables can be resolved just by implication.

This is a very simple model. However, in more complex models of the neoclassical theory, the general equilibrium will also comprise two subsystems: the real and the nominal; and the solution will also be sequential: first the real subsystem and then the nominal subsystem.

Production and distribution of equilibrium can then be represented by the following equations:

$$Y = P + W = P + w S_h \quad (3.9)$$

Total net output or national income of equilibrium (Y) is determined by quantity of employment, which in equilibrium is equal to the exogenous quantity of labor supply. Total income is distributed to capitalists as profits (P) and to workers as wages (W), which in turn is equal to the real wage rate multiplied by the quantity of employment. Income distribution of equilibrium is shown in Figure 3.1(a), where total income is equal to the area under the marginal labor productive curve, the curve HE , which is distributed to wage bill and profits.

According to the neoclassical model, this is how the market system operates. The model says that whatever the market clearing mechanism used by the markets in the real world is, it is equivalent to (not equal to) what the model says. The market solution

is the equilibrium situation in the sense that nobody has the power and the will to change the solution. It is static equilibrium in the sense that as long as the values of the exogenous variables remain fixed, the equilibrium values of the endogenous variables will be repeated period after period. If the values of the exogenous variables change, then the equilibrium situation cannot remain fixed; it will move to another equilibrium situation (assuming of course that the system is stable), which implies that the values of the endogenous variables will change as well. Then causality relations can be derived from the model.

Empirical Predictions

There are three exogenous variables in the market system: quantity of labor supplied, capital stock, and quantity of money supply. Now it is time to show how changes in these exogenous variables will change the equilibrium values of the endogenous variables. Because the interest is in the short run equilibrium, the only effect to be analyzed is changes in the money supply, maintaining fixed the stocks of capital and labor.

Suppose the government increases the money supply (say by helicopter or any similar mechanism). This increase will directly generate an excess of real cash balances willingly held by workers. They will then seek to get rid of the excess by buying a higher quantity of good B, which will tend to raise the price of good B. A higher price level would induce firms to produce more and hire more labor. Given that workers are fully employed, the competition of firms for workers would lead to a higher nominal wage rate, which would keep rising until the real wage and the employment of equilibrium are restored at the initial equilibrium. Total output would then remain constant, so will income distribution.

The price level will rise in proportion to the increase in the money supply, which will reduce the real cash balance to its original equilibrium level. The real cash balance will not change. Hence, the new equilibrium will have the real variables unchanged; the only effects of an increase in money supply are on the nominal variables: both the price level and the nominal wage rate will increase in the same proportion, just as money supply. If the money supply doubles, both nominal prices will double, leaving both the real wage rate and the real cash balance unchanged. There is no effect of money on real variables; then *money is neutral* in this model.

A final comment on the aggregation problem is in order. If money stock increases for the *individual worker alone*, we said above, he will seek to eliminate this excess by buying extra quantities of good B in the adjustment period only, which implies a negligible effect over several periods. Now it is found that if money stock increases for *everyone*, no one would be able to buy more goods even in the adjustment period. However, at the aggregate level, in their attempt to restore real cash balances, individuals end up raising the price level, and in the same proportion of money increase; as a result, the aggregate quantity demanded of good B does not change. This is a clear case of *fallacy of composition*: what is true for the individual is not true for the aggregate. Indeed, we introduced above the assumption of ignoring the real-cash balance in the individual demand for good B just by anticipating this result at the aggregate level.

3.2 The Keynesian Society

Another abstract society will now be presented. This will be called the Keynesian society. This construction is inspired in the school initiated by the English economist John Maynard Keynes (1936).

In this society not all markets will be Walrasian. In some markets, nominal prices will be exogenously determined. Some buyers or sellers will then be unable to realize the market exchange in the quantities that they are willing to exchange at the market price; so the market will operate with excess demand or excess supply. This type of market is called *non-Walrasian market*.

Economic behavior of individuals must be consistent with the context in which they operate. So, individuals need to know the range in which prices are exogenous. If the market price is above the Walrasian price, sellers will be subject to rationing. Not all sellers will be able to sell the amounts they are willing to exchange. If the market price is below the Walrasian price, buyers will face rationing. Individuals are aware of that, and their economic behavior will take this social restriction into account. The labor market is non-Walrasian and at the fixed nominal wage there is excess labor supply.

The primary assumptions of the Keynesian theory (K) can be presented as the following set of alpha propositions:

$\alpha(K).(1)$ *Institutional Context*: (a) Rules: People participating in the economic process are endowed with economic assets, which are subject to private property rights; they exchange goods under the norms of market exchange, which include the social norm that nominal wage rates cannot fall; at the initial nominal wage rate, there is excess labor supply. Walrasian and non-Walrasian markets constitute the market system; the labor market is non-Walrasian. The political regime is democratic. (b) Organizations: firms, households and the government.

$\alpha(K).(2)$ *Initial Conditions*: Factor endowments are such that the productivity of total labor is high enough to cover wages and generate profits. The exogenously given nominal wage rate is above the corresponding Walrasian nominal wage rate.

$\alpha(K).(3)$ *Economic Rationality of Agents*: Individuals act guided by the motivation of self-interest.

A Keynesian Model: The Short Run

In order to make Keynesian theory falsifiable, a Keynesian model must be constructed, which can be seen as the task of constructing a particular stage in which social actors will play their roles. The following auxiliary assumptions are introduced for that purpose:

- There are two social groups: capitalists who are endowed with stocks of physical capital and human capital, and workers with stocks of human capital and cash balances.
- Government behavior consists of supplying money to the economic process.
- Initial conditions include: factor endowments (stocks of machines and men), assets inequality among individuals, and a nominal wage rate that is set above the Walrasian price.
- Wages are paid at the end of the production period; consequently, workers need to hold cash balances for transaction purposes.
- Three markets will constitute the market system: labor, commodity, and money. Human capital is the same for all, so there is only one labor market. One sole good is produced in society, called good B. Workers and capitalist firms exchange labor services and goods in the labor and the commodity markets; money supplied by the government is used as the means of payment and is exogenously determined. There is no market for renting capital services. Capitalists and workers interact and compete in the market to obtain their objectives of self-interest, but none has the power to set prices (a perfect competitive market); they all face costless information on market prices and technology. Cost curves are different among firms.
- The economic process is static and close. Short run equilibrium is the object of analysis.

The analysis of the behavior of social actors will start with that of workers. The rationality of the typical worker consists of seeking to maximize her real income, the quantity of good B, subject to her budget constraint. Workers know that the labor market functions with job rationing. Not all workers will be able to exchange the amount of labor they are willing to sell. Once firms have determined the total labor demand D_h , at the given nominal wage rate, the quantity of labor that the capitalists will buy from each household is exogenously determined. In the labor market, there is a random rationing mechanism, exogenously determined, such that firms will buy D_{hi} units of labor from household i , where $D_{hi}=0$ for some households.

The worker i seeks to maximize real income subject to the budget constraints. Thus the set of structural equations is

$$\begin{aligned} P_h D_{hi} &= P_b D_{bi} + (D_{mi} - S_{mi}) \\ D_{mi} &= P_b L_i(D_{bi}), L_i > 0, L_i < 0 \end{aligned} \quad (3.10)$$

In this model, $D_{hi}=n_i D_h$, where n_i is exogenously given and represents the share of worker i in total employment D_h , which is the total quantity demanded by capitalists at the given nominal wage rate P_h . When $n_i=0$, the worker i is unemployed. The first budget constraint reflects the macro foundations of micro behavior, that is, the assumption that the individual worker operates in a non-Walrasian labor market. The equilibrium condition is $(S_{mi}-D_{mi})=0$. This equilibrium is (trivially) stable and then

comparative statics may be applied to derive beta propositions on the individual behavior of workers.

An increase in D_h will increase the worker's chances to get a job and thus its effect will be positive on the quantity of good B demanded. An increase in the real wage, due to changes in the nominal wage rate or the price level, will also be positive. Changes in the worker's labor supply will be ignored; changes in the aggregate labor supply will be introduced later on. A change in money endowment will increase the worker's real cash balances which he will seek to reduce by buying extra quantities of good in the period of adjustment only. After this adjustment is made, she will consume the initial quantities of good B and hold the same real cash balances, period after period; therefore, over several periods, the real cash balance effect can be ignored.

The reduced form equations are thus the following:

$$\begin{aligned} D_{bi} &= F^i(P_b, D_h, P_h), & F_1 < 0, F_2 > 0, F_3 > 0 \\ D_{mi} &= P_b L_i(D_{bi}), & L_i > 0, L_i'' < 0 \end{aligned} \quad (3.11)$$

The first equation shows the quantity of good B demanded by the individual worker and the second shows her demand for cash balances.

On the behavior of the typical capitalist, the demand for labor and the supply of good B functions will be equal to those derived for the neoclassical model. The reason is that firms face the same exogenous variables as in the neoclassical model. The only difference is that these functions cannot be expressed in terms of real wages, for the nominal wage rate is exogenous. Thus, for capitalist j, the reduced form equations are

$$\begin{aligned} D_{hj} &= H(P_b, P_h, K_{bj}), & H_1 > 0, H_2 < 0, H_3 > 0 \\ \text{Subject to } P_b S_{bj} &= P_h D_{hj} \end{aligned} \quad (3.12)$$

The first equation shows the labor demand function of firm j, expressed in terms of nominal prices. The second equation shows the budget constraint of the firm. The supply of good B is obtained from this equation.

Market General Equilibrium

Aggregating the equations in (3.11) over all workers, the following market demand functions are derived:

$$\begin{aligned} D_b &= F(P_b, D_h, P_h), & F_1 < 0, F_2 > 0, F_3 > 0 \\ D_m &= P_b L(W), & L > 0, L' < 0 \\ &= M(P_b, D_h, P_h), & M_i > 0 \end{aligned} \quad (3.13)$$

The first equation shows that the quantity of good B demanded depends on nominal prices and on the employment level. This is the result of assuming a non-Walrasian labor market, in which the income of workers do not depend on the quantity of labor they are willing to sell, but on the quantity of labor the firms are willing to buy. Although, the demand function F is homogeneous of degree zero in nominal prices, it cannot be written in terms of real wages. There is no economic meaning in the statement

“if both nominal prices double, the quantity demanded will remain unchanged,” for the nominal wage rate is fixed.

The second equation shows that the quantity of cash balance demanded depends on the price level and on the real wage bill because, in the aggregate, the quantity of good B demanded is identical to the real wage bill (W). This demand function can, in turn, be expressed in terms of the price level, the employment level and the nominal wage rate. The effects of the employment level and the nominal wage rate are both positive. The effect of the price level is more involved. A higher price level has two effects that run in opposite directions. One effect is positive because at a higher price level more cash balances will be needed; but a higher price level reduces the real wage bill and thus less cash balances will be needed. The first effect dominates over the second because it is direct and proportional, whereas the second effect is smaller due to the assumption that income elasticity is less than one; thus the net effect of the price level on the demand for nominal cash balance is positive.

Aggregating equation (3.12) over all firms, the market demand function for labor will become

$$D_h = H(P_b; P_h, K_b), H_1 > 0, H_2 < 0, H_3 > 0 \quad (3.14)$$

Subject to $P_b S_b = P_h D_h$

The quantity of labor demanded depends positively on the price level, for a higher price level implies a higher value of the marginal productivity of labor; then negatively on the nominal wage rate, for a higher nominal wage implies the need to increase the value of the marginal productivity of labor, which implies less employment (diminishing returns dictates this result); and positively on the stock of physical capital, for a higher stock implies a higher marginal productivity of labor. The constraint says that firms pay as wages what they sell in the market. Since workers spend what they earn (savings are nil), savings can come only from profits.

From these structural equations, the general equilibrium conditions of the market system can be written as

$$\begin{array}{ll} \text{Labor market} & S_h \equiv D_h + U, P_h \geq P_h^* \\ \text{Good B market} & S_b = D_b \\ \text{Money market} & S_m = D_m \\ \text{Subject to the aggregate budget constraint} & \\ & P_h D_h + P_b S_b = P_b D_b + P_h D_h + (D_m - S_m) \text{ or} \\ & P_b (D_b - S_b) + (D_m - S_m) = 0 \end{array} \quad (3.15)$$

The general equilibrium conditions include the condition that the nominal wage rate cannot be smaller than P_h^* , which is exogenously determined.

Given the values of the exogenous variables, the interaction between capitalists and workers in the market place will determine the equilibrium values of the endogenous variables of the system, which will be repeated period after period. These variables are

Exogenous variables: S_m, P_h^*, K_b, S_h

Endogenous variables: $w, D_h, U, S_b=D_b=Y, P, W, P_b, P_h$

In the aggregation, variables that were exogenous in the microeconomic equilibrium are now endogenous, such as the employment level and the price level. The market system will be able to solve this problem. The theory of markets is also applied in this Keynesian model: it assumes that the market system operates *as if* it solved a system of equations.

The core of the general equilibrium must be determined now. No subsystems exist there; hence, the solution is simultaneous. In non-Walrasian markets, the values of the quantities supplied and demanded are identical, and they cancel out in the aggregate budget constraint. Walras' Law applies to Walrasian markets only. So, of the two Walrasian markets, one is redundant. Let the commodity market be redundant. Thus, both the labor and money markets constitute the core of the system, which can solve for the two endogenous variables P_b and D_h . The core of the system includes

$$\begin{aligned} D_h &= H(P_b; P_h, K_b), H_1 > 0, H_2 < 0, H_3 > 0 \\ S_m &= M(P_b, D_h; P_h), M_1 > 0 \end{aligned} \quad (3.16)$$

The first equation shows the labor demand function, whereas the second shows the equilibrium condition in the money market. This system is simple enough to be solved graphically.

The core of the general equilibrium is represented in Figure 3.2. Panel (a) depicts the labor market, whereas panel (b) shows the money market. In order to solve for the price level, the employment level must be determined; conversely, in order to solve for the employment level, the price level must be known. The solution is simultaneous.

The other endogenous variables are determined from the core solution just by implication. Thus, the core equilibrium implies an equilibrium real wage rate (w^o), which is shown in panel (c). The market real wage rate is higher than the Walrasian wage rate (indicated by the point N). General equilibrium is with unemployment. Total output and its distribution between workers and capitalists are also determined, as can be seen in panel (c). It can also be shown there that in fact the commodity market is in equilibrium. The quantity of good demanded is equal to the wage bill; the quantity supplied is equal to total output (the area under the segment HE) minus profits, which is represented by the same area of the wage bill.

The equilibrium net output or national income (Y) and its distribution may be represented as follows:

$$Y = W + P = w D_h + P \quad (3.17)$$

Once the general equilibrium is reached, the equilibrium values of the endogenous variables will be repeated period after period as long as the exogenous variables remain fixed. This is an *equilibrium* situation, even with the existence of unemployment, because no agent has both the power and the incentive to change the situation.

Empirical Predictions

The core of the static general equilibrium is stable. The reason is that in the labor market (a non-Walrasian market) stability is trivial, as firms will be able to readjust automatically if the quantity of employment is for some reason out of equilibrium; that is, there cannot exist instability in the labor market. Therefore, the only place where instability would possibly exist is the money market (a Walrasian market). The money market is indeed stable because the demand curve is sloping downwards and the supply curve is a vertical line; hence, any price level that is for some reason out of equilibrium would restore equilibrium automatically. Stability of the general equilibrium requires stability in the money market only, which is fulfilled. Therefore, comparative statics can be applied to the general equilibrium solution to derive beta propositions.

The exogenous variables include the capital stock, the quantity supplied of workers, the stock of money, and the nominal wage rate. The short run model will show the effects of changes in the money supply and the nominal wage rate upon the main endogenous variables, total output and its distribution.

If the stock of money increases, the direct effect will be to increase both the price level and the employment level. The reason is that workers will try to get rid of the excess cash balances by buying more quantities of output, which will tend to increase the price level. Firms would then have incentives to produce more and hire more workers, which is viable because there is unemployment. But that will not be all; some additional effects may take place. The real wage rate would fall, which would then tend to change the real wage bill, which in turn would change the demand for money. However, the change in the real wage rate is ambiguous and would have a small effect or none upon the demand for money and on the price level; then the initial direct effect will prevail.

This effect can be seen in Figure 3.2 (c). An increase in money supply raises the price level, which in turn reduces the real wage rate, which implies a rise in the employment level. So the initial equilibrium situation, at point E, will move to another equilibrium situation, say to a point E' that lies below point E along the labor demand curve HN. Unemployment is thus reduced. Total output increases. The change in income inequality is ambiguous because the change in the real wage bill is ambiguous.

This Keynesian model also predicts that a sufficiently large increase in the supply of money will ultimately result in full employment in the labor market. Full employment can be attained through government policies. If money supply is increased further, the initial effect will be to increase the price level, which induces firms to hire more labor; but there are no unemployed workers to hire; hence firms will compete for workers and the nominal wage rate will rise. The ultimate effect of an increase in money supply will be to increase the price level and the nominal wage rate in the same proportion, leaving unchanged the real wage rate. Under full employment, money becomes neutral, as in the neoclassical model.

The effect of an exogenous increase in the nominal wage rate will directly increase the real wage rate, which will lead firms to reduce the quantity demanded for labor. Firms will tend to reduce employment and thus total output will tend to fall. The wage bill change is ambiguous; therefore, the demand for money and the additional effect upon the price level will also be ambiguous; hence, the direct effect will prevail.

In Figure 3.2 (c), the initial equilibrium situation at point E will move to another equilibrium situation, say to a point E' that lies above point E, along the labor demand curve. Unemployment increases. Total output falls. Change in income inequality is ambiguous.

The labor market in this Keynesian model works as follows. Whenever there is an excess labor demand situation, the nominal wage will rise (as in a Walrasian market); but whenever there is an excess labor supply situation, the nominal wage cannot fall, and the situation will remain unchanged as the equilibrium situation. Nominal wages are sticky downwards, but not upwards.

The Keynesian model presented here is a very simple one. Macroeconomic textbooks present more complex Keynesian models in which the bond market is included. The consequence is that the government increases money by buying bonds from the public at a price that clears the market; moreover, the bond price and the interest rate are inversely related, that is, if the bond price rises, the interest rate falls. This way of increasing money is called "open market operation," while the mechanism utilized in the model presented above is known as "helicopter money," for money is injected directly in the economy as if by using a helicopter. Therefore, in the textbook model, the government can choose as policy instrument either money supply, nominal exchange rate, or nominal interest rate; that is, there is only one degree of freedom: once one of them is chosen, the other two will be determined endogenously. The essential relations between nominal and real variables are thus captured in the simpler model presented here.

3.3 The Classical Society

The classical society will also be presented as an abstract society. This construction is inspired by the school initiated by David Ricardo (1821) and Karl Marx (1867).

Property structure is such that the classical society is a class society. There are two social classes: capitalists who are endowed with physical capital and workers who are endowed with human capital alone. Capitalists do not rent out the services of their capital in the market because they wish to avoid the formation of new firms and new capitalists. Firms can only be created by accumulating physical capital.

In the classical society, the labor market operates with the real wage rate of subsistence (w^*), which satisfies the worker's subsistence needs and is exogenously given. The theory assumes that at the subsistence real wage rate, there is excess labor supply. The labor market is thus non-Walrasian; the price of labor services cannot clear the labor market.

Because capitalists control the ownership of physical capital, capitalists are able to set the length of the working time at a level that is above of what would be needed to produce the quantity of subsistence goods for workers. This length is a parameter of the production function. Thus, the total work length utilized by the firm (call it η) has two components: the necessary labor (η^*), which is that part of the length of the working time that would be enough to produce the quantity of workers' subsistence goods; and the surplus labor (η^e), which is that part that produces surplus output. Capitalists

appropriate the output produced with the surplus labor in the form of profits. This appropriation is called exploitation. Profits come from exploitation. Without surplus labor, profits could not exist.

To illustrate these concepts, suppose a firm produces 100 kg of output per worker in a week, with 40 hours as the length of working time; also suppose the subsistence income per worker is equal to 50 kg per week. The production process can be conceptually separated into “necessary labor”, which is equal to 20 hours (the time necessary to produce 50 kg for the subsistence of the worker), and the “surplus labor”, the other 20 hours, which will produce the other 50 kilos. The necessary labor gives rise to wages, whereas the surplus labor generates profits.

This example shows that labor productivity must be sufficiently high to generate output per worker beyond subsistence wage. If labor productivity were 50 kg, this society could not generate profits and could not operate as a capitalist society. If labor productivity were less than 50 kg, this society could not be economically viable, not only under capitalism, but under any type of social organization; it could not exist.

The primary assumptions of the classical theory (CL) can be expressed as the following set of alpha propositions:

$\alpha(CL).(1)$ *Institutional context*: (a) Rules: People participating in the economic process are endowed with economic assets, which are subject to private property rights; people exchange goods subject to the norms of market exchange, which includes the norm that market real wages cannot be smaller than the cost of the worker’s subsistence basket of goods. The market system operates with Walrasian and non-Walrasian markets; the labor market is non-Walrasian. (b) Organizations: Firms, households, and the government.

$\alpha(CL).(2)$ *Initial conditions*: Individuals are endowed with unequal quantities of economic assets. A social group concentrates the total stock of physical capital; there are two social classes: workers and capitalists.¹ At the initial real wage rate, which is exogenously determined, there is excess labor supply.

$\alpha(CL).(3)$ *Economic rationality of agents*: Individuals act guided by the motivation of self-interest. Capitalists seek two particular objectives: profit maximization and the preservation of their class position. Capitalists do not rent out the services of their capital stocks in the market for they prefer profits to rents. Capitalists seek to set the length of the working day above the socially necessary working time.

A Classical Model: The Short Run

In order to generate beta propositions, a classical model is constructed with the introduction of consistent auxiliary assumptions. They are:

¹ While the assumption of two social classes in the neoclassical and Keynesian models above was introduced as an auxiliary assumption to generate a model, this class structure is a primary assumption of the classical theory.

- There are two social groups: capitalists who are endowed with stocks of physical capital and human capital, and workers with stocks of human capital and cash balances. Government behavior consists of supplying money to the economic process.
- Wages are paid at the end of the production period; consequently, workers need to hold cash balances for transaction purposes.
- Three markets will constitute the market system: labor, commodity, and money. Human capital is the same for all, so there is only one labor market. One sole good is produced in society, called good B. Workers and capitalist firms exchange labor services and goods in the labor and the commodity markets; money supplied by the government is used as the means of payment and is exogenously determined. Capitalists and workers interact and compete in the market to obtain their objectives of self-interest, but none has the power to set prices (a perfect competitive market); they all face costless information on market prices and technology. Cost curves are different among firms.
- The economic process is static and close. Short run equilibrium is the object of analysis.

The study of the general equilibrium starts with the behavior of workers. Workers exchange their labor power in a labor market in which not all workers can get the amount of jobs they are willing to exchange at the prevailing real wage rate. In the labor market, there is a rationing mechanism, which is assumed to be random. Firms, once they know the quantity of workers they want to hire D_h , will buy D_{hi} units of labor from each household i , where $D_{hi} = 0$ for some households.

In this context, the typical worker i seeks to maximize his real income, subject to the following structural equations

$$\begin{aligned} w D_{hi} &= D_{bi} + (D_{mi} / P_b - S_{mi} / P_b) \\ D_{mi} / P_b &= L_i(D_{bi}), L_i > 0, L_i < 0 \\ (S_{mi} - D_{mi}) / P_b &= 0 \end{aligned} \quad (3.18)$$

Here $D_{hi} = n_i D_h$, where n_i is the share of worker i in total labor demand D_h . When $n_i = 0$, the worker is unemployed. The first equation of the budget constraint reflects the macro foundations of micro behavior: the worker operates in a context of non-Walrasian labor market and Walrasian commodity and money markets. The real income of the individual worker depends on the quantity bought by firms, not on the quantity sold by the worker (as is the case in the Walrasian labor market). The second equation shows the real cash balance constraint. The third equation is the equilibrium condition.

From the structural equations shown above, it follows that the equilibrium condition is $(D_{mi} - S_{mi}) = 0$. This condition is (trivially) stable. Comparative statics may now be applied to derive beta propositions: the effect of the exogenous variables $(w, D_h, P_h, P_b, S_{mi})$ upon the endogenous variables (D_{bi}, D_{mi}) .

An increase in the stock of money endowments will lead the worker to get rid of the extra money by buying extra quantities of goods to restore his equilibrium real cash

balance. As in the neoclassical model, it is assumed that the real cash balance effect at the individual level is negligible and can safely be ignored. Changes in labor supply are relevant at the aggregate level only. The effects of the other exogenous variables are similar to those obtained in the neoclassical model. The individual demand functions for good B and for money can then be written as the following equations:

$$\begin{aligned} D_{bi} &= F^i(w, D_h), F_{11}^i > 0, F_{22}^i > 0 \\ D_{mi} &= P_b L_i(D_{bi}), L_{i1} > 0, L_{i2} < 0 \end{aligned} \quad (3.19)$$

On the other hand, the individual firm j seeks to maximize profits, subject to its individual endowments of capital and production function. For firm j we have

$$\begin{aligned} \text{Maximize } P_j' &= P_b Q_{bj} - P_h D_{hj} \\ \text{Subject to } Q_{bj} &= \eta q_{bj} = \eta \phi_j(D_{hj}, K_{bj}) \end{aligned} \quad (3.20)$$

The production function of the firm refers to output q_b , which represents the output in the unitary period of production (say the hour). This unitary period can be replicated for η periods (say, 40 hours per week) and the firm will be able to produce proportionally more output; that is, the production function is homogeneous of degree one with respect to time. The work length or duration of the economic process η is exogenously given and is subject to $\eta \leq \eta'$, where η' is the maximum duration that is physically viable for workers (say, 60 hours per week). Hence, Q_{bj} is the flow of output (per week).

From the structural equations shown above, it follows that the firm's equilibrium condition is that the marginal productivity of labor should be equal to the real wage rate; that is

$$\eta \phi_{j1}(D_{hj}, K_{bj}) = w \quad (3.21)$$

Profits are maximized when the marginal productivity of labor (the additional output per week that results from an additional worker) must be equal to the weekly real wage rate.

Because the wage rate must be equal to the subsistence wage w^* , which is equal to the output generated with the necessary labor, the following relations hold true for the typical firm j :

$$w = w^* = \eta^* (q/D_h)_j = \eta \phi_{j1}(D_{hj}, K_{bj}) < \eta (q/D_h)_j \quad (3.22)$$

The first equality indicates that the market wage rate must be equal to the subsistence wage. The second equality indicates that, given the hourly average productivity of labor and given the subsistence wage rate, the necessary labor η^* becomes determined endogenously (say 20 hours per week). The third equality shows just the profit maximization condition, stated above; and the last inequality shows that the marginal productivity of labor is smaller than the average productivity of labor, which follows from the assumption of diminishing returns.

The set of relations shown in (3.22) implies that $\eta > \eta^*$ (say η is equal to 40 hours per week). Because $\eta = \eta^* + \eta^e$, this condition in turn implies that $\eta^e > 0$. Hence, if there is no surplus labor, there will be no profits. This set of relations also says that positive profits imply that the weekly average productivity of labor must be higher than the weekly marginal productivity of labor. The labor exploitation account is therefore equivalent to the labor productivity account in explaining the generation of profits.

The firm's equilibrium is (trivially) stable. Comparative statics can then be applied to derive beta propositions on the firm's behavior: the effect of changes in the values of the relevant exogenous variables (w , K_{bj}) upon the relevant endogenous variables (D_{hj} , P_j).

The quantity of labor demanded depends negatively upon the real wage rate and positively upon the stock of physical capital. The supply of good B must be equal to the real wage bill the firm must pay for the hired labor, as a condition of its budget constraint. Hence the reduced form of the behavior of firm j can be written as

$$\begin{aligned} D_{hj} &= H^j(w, K_{bj}), H_1 < 0, H_2 > 0 \\ \text{Subject to the budget constraint } S_{bj} &= Q_{bj} - P_j = w D_{hj} \end{aligned} \quad (3.23)$$

General Equilibrium

In order to construct a general equilibrium system, the structural equations showing market behavior of the social actors are needed. Aggregating over all workers, the system (3.19) becomes

$$\begin{aligned} D_b &= F(w, D_h), F_1 > 0, F_2 > 0 \\ D_m &= P_b L(D_b), L > 0, L' < 0 \\ \text{Subject to the budget constraint: } w D_h &= D_b + (D_m / P_b - S_m / P_b) \end{aligned} \quad (3.24)$$

The particular feature of this model is that the quantity of good B demanded depends positively upon both the real wage rate and the quantity of labor demanded (not the quantity of labor supplied, as in the neoclassical model). The higher the employment level D_h , the higher the demand for good B.

Aggregating over all firms, equation (3.23) becomes

$$\begin{aligned} D_h &= H(w, K_b), H_1 < 0, H_2 > 0 \\ \text{Subject to the budget constraint: } S_b &= w D_h \end{aligned} \quad (3.25)$$

This is the aggregate labor demand function and the constraint shows the aggregate supply of good B, which is obtained directly from the aggregate budget constraint of firms.

The equilibrium conditions for general equilibrium are

$$\begin{aligned} \text{Labor market} \quad S_h &\equiv D_h + U, w \geq w^* \end{aligned} \quad (3.26)$$

$$\begin{aligned}
&\text{Commodity market} \quad S_b = D_b \\
&\text{Money market} \quad S_m = D_m \\
&\text{Subject to the aggregate budget constraint:} \\
&w D_h + S_b = D_b + (D_m / P_b - S_m / P_b) + w D_h \text{ or} \\
&(D_b - S_b) + (D_m / P_b - S_m / P_b) = 0
\end{aligned}$$

The size of unemployment (U) is now a *possible* outcome of the economic process because the market real wage cannot be smaller than the subsistence wage w^* .

Given the values of the exogenous variables, the interaction of workers and capitalists in the market system will determine the values of the endogenous variables of the system. These variables are

$$\begin{aligned}
&\text{Exogenous variables: } K_b, S_h, S_m, w^*, \eta \\
&\text{Endogenous variables: } D_h, w, U, Q_b, P, W, P/W, P_b, P_h
\end{aligned}$$

In order to determine the core of the model, take note that two Walrasian markets and one non-Walrasian market constitute the market system. Among Walrasian markets, one market is redundant; let this redundant market be the good B market. In addition, two subsystems can be found, a real (labor and commodity markets) and a nominal (money market); moreover, the monetary subsystem needs a real variable to reach a solution, but not vice versa. Thus, the general equilibrium solution can be found by simply solving for the labor market and the money market, and solving them sequentially. Thus the core of the general equilibrium system is composed of the labor market alone, which will solve for the employment level. The money market can then solve for the price level. The rest of the endogenous variables will be solved just by implications.

This classical model is sufficiently simple to find the solution graphically. The labor market is presented in Figure 3.3, panel (a). The market real wage rate is equal to the subsistence wage rate (w^*), at which there exists excess labor supply. The employment level is determined by the labor demand curve HN. Point E shows the equilibrium situation in the labor market. (The Walrasian real wage rate—at point N—is below the subsistence wage rate.) Total wage bill is then determined. Total profit is equal to output produced and not paid as wages, the result of surplus labor.²

The money market is shown in Figure 3.3, panel (b). Once the wage bill is known, the demand for money is also known, the curve L. Given the quantity of money supplied, the price level is determined by the money demand curve. So, sequentially, equilibrium in the money market is determined, once the equilibrium in the labor market is solved.

² Roemer (1982) has developed a general theory of exploitation where surplus labor can be generated via the labor market or the credit market. In the latter case, surplus labor must generate not only profits but also interest payments that go to banks. This is theoretically correct; empirically, however, in this model profits are considered the most important component of surplus labor. In the real world, profit incomes are indeed much larger than interest incomes in national income accounts.

The general equilibrium solution is with unemployment. Can this be an equilibrium situation? The classical theory had the intention to show the shortcomings of the capitalist system. This is a class conflict society, where workers if hired are exploited, but if not exploited are unemployed and have zero wage income.

But the unemployment situation will generate social tensions. Marx predicted the breakdown of the capitalist system because of this and other social tensions created in the economic process, especially by the calvaries of capitalism observed in its initial periods. This prediction, however, has been refuted by reality. A reason may be found in the introduction of institutional innovations in most of the First World countries, such as the unemployment insurance, which transfers incomes to the unemployment through the public budget. The amount of this transfer should be lower than the market wage rate so as to maintain the incentives for both employment seeking and labor discipline. Introducing the assumption that unemployment insurance exists, which gives the unemployed a transfer of income that is lower than the market wage rate, general equilibrium with unemployment is obtained. This situation can be repeated period after period because no social agent has the power and the will to change it.

From the solution of the core, other endogenous variables can be solved by implication. It can easily be verified that the commodity market is also in equilibrium: the area showing the quantity demanded (the wage bill) is the same area showing the quantity supplied (total output minus profits). Once the price level is known, the nominal wage is endogenously determined, for it must be consistent with the exogenous real wage rate.

The net output that comes out of the production process is called the *economic surplus* (ES). According to the classical model presented here, workers get only subsistence wage in the equilibrium situation. Hence, capitalists appropriate the entire economic surplus in the form of profits. Thus,

$$ES = P = \kappa K_b \quad (3.27)$$

Note that κ is not a market price, but the average rate of return of capital.

Once the values of the endogenous variables are determined, these values will prevail period after period, as long as the exogenous variables remain fixed. This is an equilibrium situation because no one has both the power and the incentive to change the situation.

Empirical Predictions

In this classical model, the general equilibrium is stable. Because the solution is sequential, it is sufficient to show that there exists stability in the equilibrium in the labor market, the core of the system, and in the money market, taken separately. That this is indeed the case can be seen in Figure 3.3. The comparative statics method can then be applied to derive beta propositions. For the short run, the relevant exogenous variables are reduced to the money supply.

An increase in the aggregate money supply will affect the equilibrium real cash balances in the money market. Workers would seek to get rid of the excess cash balances by buying extra quantities of good B in the period of adjustment. Then the price level would tend to increase and thus firms would tend to produce more output and hire more labor, which could be materialized because there is unemployment. However, the subsistence real wage would tend to diminish, which by assumption is not admitted. Then the nominal price would have to increase and do so until the subsistence real wage is restored.

In the new equilibrium, all real variables in the labor market will then remain unchanged. The only effect of a change in the money supply is to change the nominal variables; it does not affect the real variables; hence, *money is neutral*.

In the classical society, the strict problem of distribution refers only to the distribution of surplus. In this particular model the entire surplus goes to profits. More generally, the theory allows for models in which the economic surplus can be appropriated partly by capitalists as profits and partly by workers as higher-than-subsistence wage. Workers may appropriate part of the economic surplus depending on their relative strength in the distribution conflict. Thus, under classical theory not all the observed wage bill is part of the economic surplus; it includes only that part that exceeds the subsistence wage. Subsistence wage is the cost of maintaining workers, similar to the cost of maintaining horses or physical capital goods. It could not be part of the net output.

In this classical model, however, the category “national income” will also be used to discuss income distribution outcomes. The reason is that the wage bill and profits are both endogenous. The conclusion is, however, that the effect of exogenous variables upon distribution of national income between workers and capitalists is ambiguous.

Finally, this classical model shows that a continuous exogenous increase in the stock of physical capital will shift continuously the labor demand curve and will eventually eliminate the excess labor supply. Then there will eventually be excess demand for labor and the competition among firms for workers will imply equilibrium wage rates that are higher than the subsistence wage; hence the real wage rate will endogenously rise. The same effect will result if the quantity of labor supplied decreases continuously. Under sufficient large changes in these exogenous variables, therefore, the labor market will become a Walrasian market, which means that this classical model may become the neoclassical model studied earlier.

3.4 Empirical Consistency: The First World Countries

From the list of eight empirical regularities about capitalism show in Chapter 2, two refer to the First World behavior in the short run. They are the existence and persistence of unemployment (Fact 1) and the interaction between nominal and real variables (Fact 4). Do the three theories presented here explain these facts?

The empirical predictions of the neoclassical model are refuted by Facts 1 and 4. The empirical predictions say that the labor market equilibrium is with full employment and that money is neutral.

The predictions of the Keynesian model are consistent with Fact 4. With respect to Fact 1, there are some difficulties. This model can predict the existence of equilibrium with unemployment; however, there are two problems with the Keynesian model. First, there is a logical problem with the existence of unemployment. The nominal wage rate is not truly exogenous, for it must be set above the Walrasian wage rate if it is to generate unemployment initially. Second, the model also predicts equilibrium with full employment in the short run. When would the model be refuted? Never. The existence of unemployment cannot refute the theory; the inexistence of unemployment could not either. The model is not falsifiable because it predicts all possible outcomes. (It is similar to the tautological statement “It will rain or not rain here tomorrow.”) The bottom line is that in the Keynesian model unemployment is not a necessity; unemployment plays no role in the functioning of capitalism. In sum, the Keynesian model has difficulties in explaining Fact 1.

Finally, Fact 4 refutes the classical model. Money is neutral in this model. Regarding Fact 1, the classical model faces the same type of difficulties shown above for the Keynesian model. First, the existence of unemployment is exogenously determined. Second, the model also predicts full employment equilibrium in the long run. Other models have introduced mechanisms to maintain the “industrial reserve army” (unemployment) in the process of capital accumulation, so as to *avoid the increase in the real wage rate in the long run*, by assuming endogenous technological change that is labor-saving, which displaces labor continuously. However, this prediction is refuted by Fact 5.

3.5 Generalizing the Models to Theories

The neoclassical, Keynesian, and classical models that we have presented in this chapter show difficulties in explaining the basic short-run facts of capitalism. But these models are certainly very elementary. One would tend to believe that other models that are constructed under different assumptions, such as different markets structures or assuming an open economy, would be needed in order to draw more definitive conclusions about the explanatory power of these theories.

However, it is easy to see that *any* neoclassical model will predict labor market equilibrium with full employment; it will also predict that money is neutral. The reason is that any neoclassical model will assume Walrasian markets, including Walrasian labor market. An implication of this assumption is that nominal prices will play the role of clearing the markets, including the nominal wage rate in the labor market. The other implication is that real endogenous variables will be determined by real exogenous variables alone, whereas exogenous nominal variables will affect only nominal variables.

But then, how does neoclassical theory explain the existence and persistence of unemployment? The observed unemployment is considered just *frictional*. The model presented above has assumed that workers and jobs are homogeneous. Suppose now

that workers and jobs are heterogeneous, which implies that worker and job matching will take time; also suppose that there exists a rate of job separations by firms in the period of analysis. The job matching process implies that at any period of time there would be workers seeking jobs and also firms seeking workers. Thus unemployment would be the result of the process of job searching by workers under imperfect information. Then it follows that, given the conditions of demand for labor and supply of labor, there would exist a Walrasian wage rate in the labor market; but the job matching process would lead to the existence of a “natural rate of unemployment,” at which the rate of finding jobs is equal to the rate of job separations (Barro 1997, Chapter 10).

This argument is logically inconsistent with the assumption of Walrasian labor market, in which the nature of the market is such that the nominal wage rate would change until the market clears. But under equilibrium with frictional unemployment, the nominal wage rate plays no role in the equilibrium adjustment, which occurs via quantity adjustments. Frictional unemployment would be consistent with the Walrasian market assumption if the mechanism of quantity adjustment would lead toward equilibrium with zero (or negligible) unemployment; that is, as the economic process is repeated period after period, then frictional unemployment would have to decline over time towards zero, not towards a natural rate of unemployment. Neoclassical theory cannot explain unemployment; the labor market does not operate as a Walrasian market.

Similarly, the conclusion of the particular classical model shown here can be extended to the entire family of short run models of the classical theory. The reason is that any classical model will assume an exogenous real wage rate. Therefore, in the short run, unemployment may or may not exist, depending on the factor endowments of society. Any classical short run model will also predict money neutrality, which is derived from the assumption of exogenous real wage rate.

In the case of the Keynesian model, it is also simple to show that its results can be generalized to the entire family of models. Any Keynesian model will assume “exogenously” determined nominal wage rates, which is set to generate excess labor supply. Therefore, any model will predict both equilibrium with unemployment and equilibrium with full employment. In fact, Keynesian theory was created to reach full employment through public policies. Hence, unemployment plays no role in the functioning of capitalism, which can operate with full employment as well.

The basic conclusion of this chapter is that three basic economic theories are refuted by facts of short run behavior of First World capitalism. Fact (1) refutes the three theories, whereas Fact (4) refutes neoclassical and classical theories. An economic theory that seeks to explain the existence and persistence of unemployment—to explain why capitalism *always* operates with unemployment—needs to show that unemployment plays a role in the functioning of capitalism, in the short run and in the long run. A new economic theory of the First World that seeks to explain Fact 1 and Fact 4 will be presented in the next chapter.

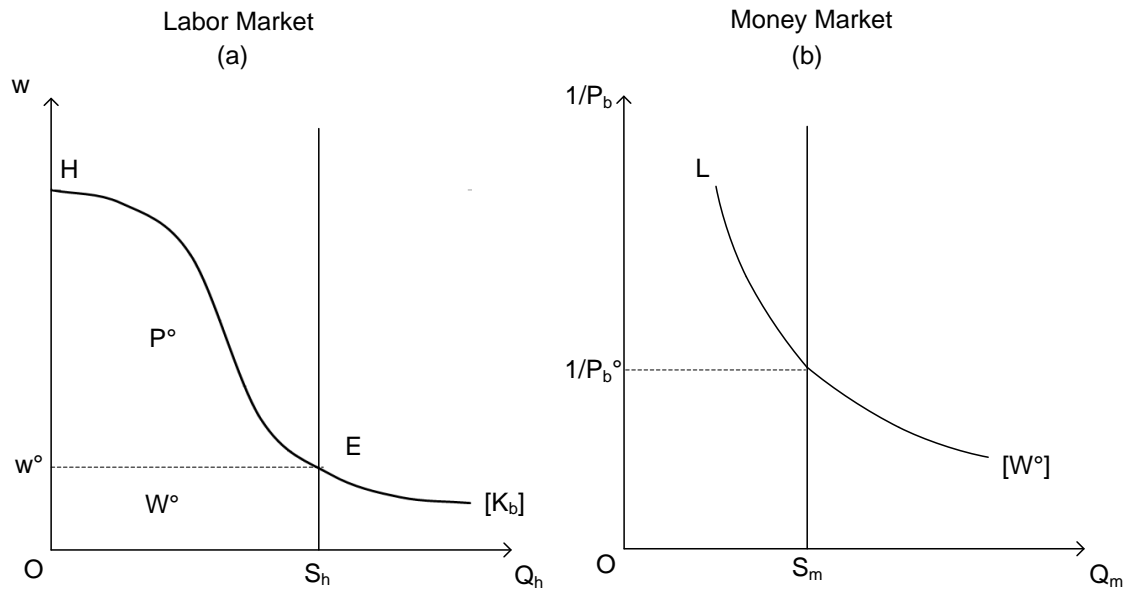
Figure 3.1. Neoclassical General Equilibrium

Figure 3.2. Keynesian General Equilibrium: Labor and Money Markets

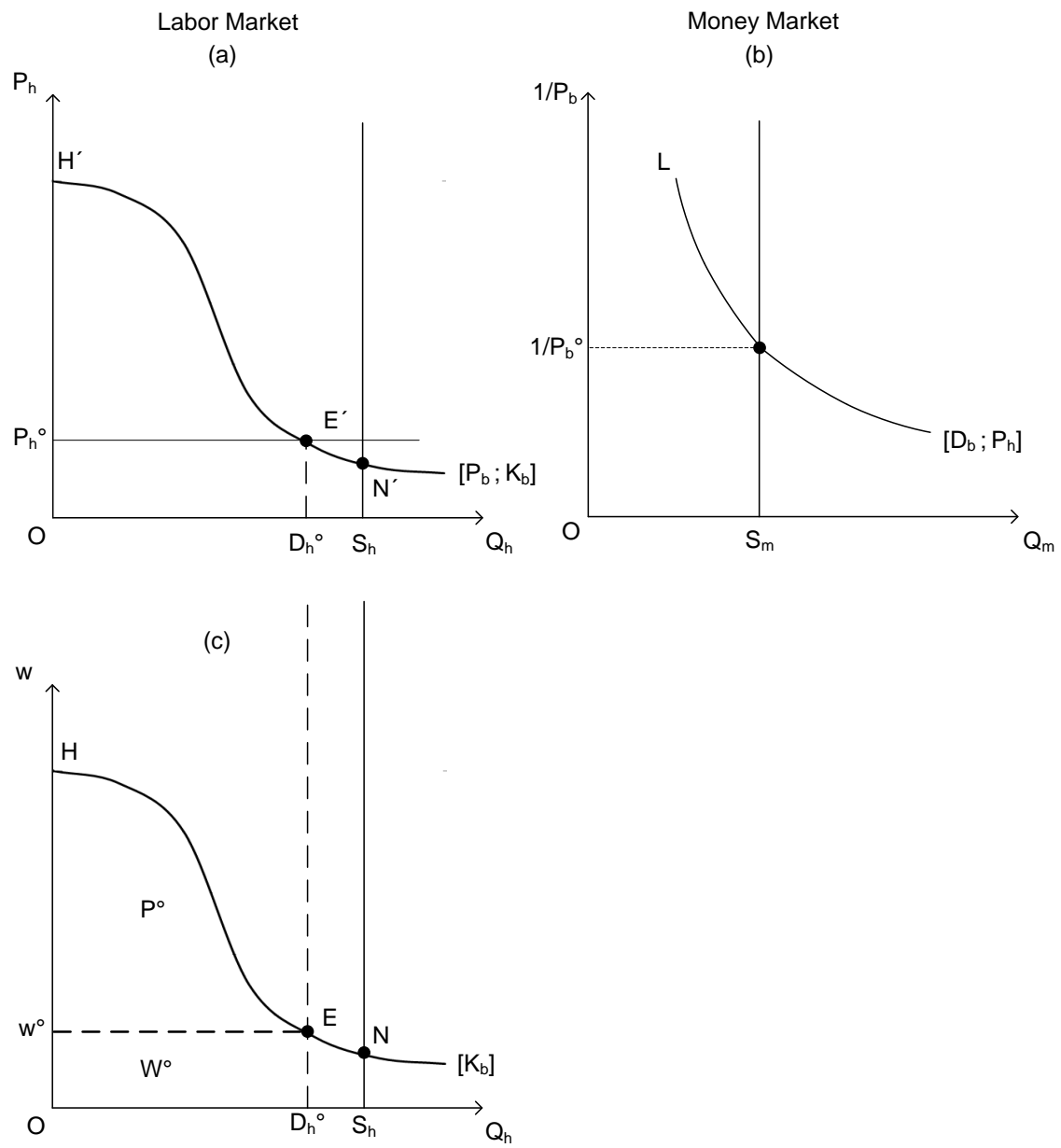
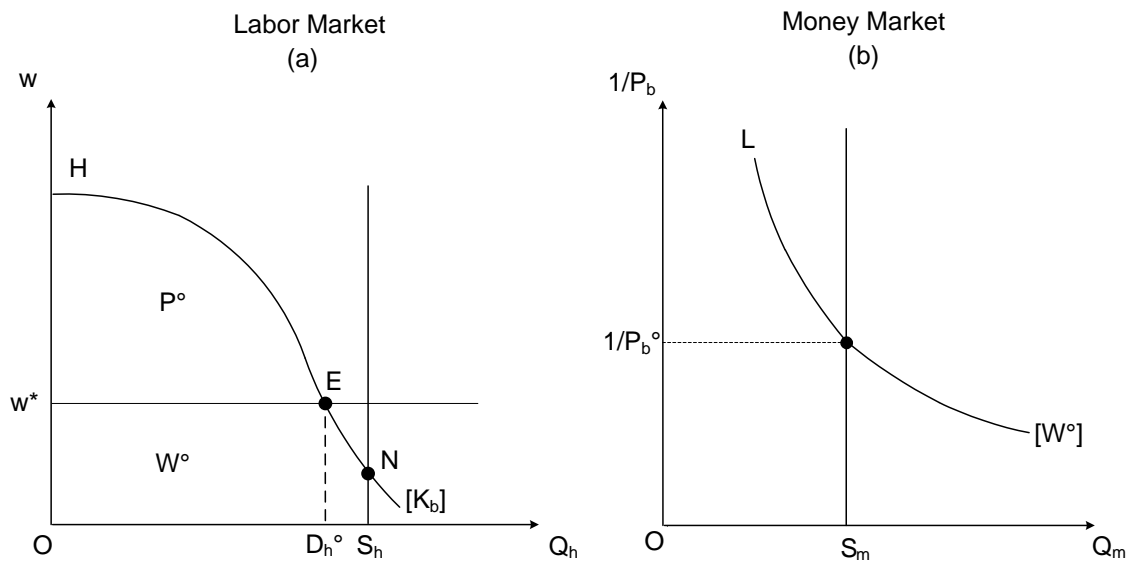


Figure 3.3. Classical General Equilibrium



PART I

**THREE TYPES OF CAPITALIST SOCIETIES:
PARTIAL THEORIES**

CHAPTER 4

THE FIRST WORLD

Any economic theory that attempts to explain production and distribution in the First World countries must construct an abstract society in which two outcomes of the economic process are necessary. First, the labor market must show that unemployment is a necessity for the functioning of this society. Second, there must be interplay between real and monetary variables. These constitute notable empirical regularities of the First World countries (Facts 1 and 4, as listed in Chapter 2). A new abstract society—called epsilon society—will be developed now with the intention to meet those challenges.

4.1 Epsilon: Socially Homogeneous Class Society and Under-populated

Epsilon is an abstract capitalist society. It is a class society, in which two social classes exist: capitalists and workers. This class division of society is the result of the unequal distribution of economic assets between individuals. Capitalists concentrate the property of physical capital. This is an assumption taken from the classical school, of which David Ricardo (1821) and Karl Marx (1867) are the founding fathers.

Another assumption about the initial conditions of epsilon society refers to factor endowments, that is, capital per worker. The assumption is that the average productivity and marginal productivity of the total labor force in society depends positively upon the capital per worker endowment. The marginal productivity of the total labor force could be either positive or zero. When it is positive, we say that society is *under-populated*. The assumption is that epsilon is an under-populated society.

On the social institutions under which the interactions between social classes operate in epsilon, the assumption is that individual freedom is the bases of social relations. Therefore, the market system and the democratic political system constitute the basic institutions of capitalism in general and of epsilon in particular. Exchange of goods takes place under the form of market exchange, which includes the norms of private property of capital and voluntary exchange. This assumption is taken from Adam Smith (1776). The democratic political system includes the norm that individuals are entitled with the same political rights and duties; thus, epsilon is a socially homogeneous society.

Labor market institutions are also established. Labor exchange operates under the norms of market exchange, which includes free labor and voluntary exchange (any form of slavery is not allowed). Another norm is that it is considered socially unfair to reduce nominal wages. The reduction of real wages may occur only via increases in the price level and inflation. The nominal wage rate may go up if firms find it profitable to

do so; its stickiness is downwards, not upwards. This assumption is taken from John Maynard Keynes (1936).

Individuals act guided by the motivation of self-interest. Workers and capitalists follow different objectives. Capitalists want to extract as much effort as possible from workers in order to maximize profits. Workers want to earn the highest real wage with the lowest effort. This social conflict in labor relations arises because of workers' exclusion from the property of the firm's capital stock. Therefore, capitalists need to use particular devices to extract effort from workers. These assumptions are taken from Bowles (1985), Kalecki (1971), and Shapiro and Stiglitz (1984).

Another primary assumption is that capitalists seek two objectives that are hierarchically ordered. Capitalists seek, firstly, to remain as members of the capitalist class and, secondly, profit maximization. Therefore, secure property rights has priority over profits; that is, there is no substitution between profits and the risk of losing property rights or losing capital endowments.

Finally, on organizations, the social institution includes firms, households, and the government. The government is just another social actor that interacts with capitalists and workers; it is not an actor that is above social classes.

The social norms constitute formal rules of the game that set constraints upon individual behavior and social interactions. Whether those rules will operate fully or partially under the individual behavior that is based on selfish rationality is something that the theory needs to explain.

The primary assumptions of the epsilon theory (ϵ) can be summarized as a set of alpha propositions as follows:

$\alpha(\epsilon).(1)$ *Institutional Context*: (a) Rules: People participating in the economic process are endowed with economic and political assets. Economic assets are subject to private property rights. People exchange goods subject to the norms of market exchange, which include the particular norm that in labor markets nominal wages cannot fall. The political rights and duties are uniformly entitled; so there is only one class of citizenship in the democratic system. (b) Organizations: households, firms, and the government.

$\alpha(\epsilon).(2)$ *Initial Conditions*: (a) Initial inequality: individuals are endowed with unequal quantities of economic assets, but equal political entitlements; hence, there are two social classes: capitalists and workers. (b) Initial factor endowments: the stock of capital per worker is such that the marginal productivity of the entire labor force is largely positive.

$\alpha(\epsilon).(3)$ *Economic Rationality of Agents*: Consistent with the institutional context, individuals act guided by the motivation of self-interest. Capitalists seek two particular objectives, hierarchically ordered: firstly, maintenance of class position and secondly maximization of profits. In the labor market, workers seek to maximize wages and minimize effort, while capitalists seek to minimize wages and maximize effort. Due to this conflict in labor relations, capitalists use particular devices to extract effort from workers.

4.2 The Nature of the Labor Market

Consistent with the institutional context of epsilon society, the existence of social classes implies conflictive labor relations, which in turn imply that the labor market cannot operate as a Walrasian market. If all workers willing to exchange labor were able to find jobs at the prevailing wage rate, would there be incentives for workers to supply their highest work effort to the firm? What if workers are found shirking and are then dismissed? Because the labor market is Walrasian, the dismissed workers will always be able to find jobs at the prevailing market wage rate. What would then be the cost of shirking for workers? Zero cost. Workers would then have no incentives to provide their highest work effort in the working place.

When workers and capitalists are not full partners, capitalists must use incentive devices to extract work effort from workers. The incentive device must make shirking costly. What device would do that?

Consider a model in which unemployment is the device that capitalists use to discipline workers. Full employment is against the interest of capitalists; under full employment, unemployment would cease to play its role as a disciplinary device, as the Polish economist Michael Kalecki argued long time ago (Kalecki, 1971, chap. 12). More recent contributions are due to Shapiro and Stiglitz (1984) and Bowles (1985).

In order to create incentives for work effort, an individual firm will seek to pay a premium and offer a higher wage rate than the prevailing Walrasian wage rate. Then dismissed workers will suffer a cost, the lower wage that prevails in the market. But the other firms will also follow this procedure and thus the average wage rate will rise above the Walrasian wage rate. But then the individual firm, acting guided by the motivation of self-interest, will again increase the wage rate. The other firms will also follow this procedure, and so on.

In order to find a wage rate of convergence, the model will assume that the premium rate, as percentage of the market wage rate, is uniform for all firms and it decreases in every round; thus a final wage rate of equilibrium will be reached. The final outcome is that the new market wage rate will lie above the Walrasian price. The equilibrium nominal wage rate implies excess labor supply or unemployment. Then dismissed workers will not move from one firm to another, but most likely will move to unemployment. Shirking behavior will be costly for the worker. The aggregate incentive device to discipline workers is unemployment.

Unemployment is the device used by capitalist firms to secure labor discipline. Unemployment now plays a significant role in the functioning of the labor market. Assume that a minimum unemployment rate (call it u^*) is needed to extract work effort from workers. The observed unemployment rate could be higher, due to other factors.

If there must be unemployment in the labor market, then the wage rate cannot be equal to the Walrasian price, but it will be higher than that. Moreover, given the labor market conditions of demand and supply, there is a relation between real wage rate and unemployment rate. When unemployment is zero, we know that the market price of labor is equal to the Walrasian price; therefore, a given unemployment rate implies a

real wage that is above the Walrasian, and vice versa. The wage rate that corresponds to the minimum unemployment rate to generate efficiency in labor productivity may be called *the minimum efficiency wage rate* (w^*). Therefore, wages rates in the labor market must be equal to or higher than the efficiency wage rate as a device to secure unemployment and thus labor discipline, which will in turn ensure high labor productivity levels and profit maximization in the firms.

The unemployed would be willing to be hired at the prevailing market wage or even at a lower wage rate; nonetheless, they cannot bid the wage rate down as a way to get employed. Firms would not accept that because they know that if they hired workers at lower wage rates, work effort and labor productivity would decrease. Thus no one has the power and the incentive to change this solution and there is equilibrium with unemployment in the labor market. The labor market is non-Walrasian. This is a case in which, notwithstanding the endogenous determination of the real wage rate, the labor market is non-Walrasian. The equilibrium real wage is not the Walrasian price and the labor market operates with excess labor supply.

In sum, in epsilon society, the labor market operates with unemployment; it is a non-Walrasian market. A proportion of workers will be excluded from the labor market. The labor market could not operate as a potato market (the paradigm of a Walrasian market); potatoes cannot change their behavior depending on their market price or depending on what the excess supply is, but workers can. This is the nature of the labor market.

This simple model can be presented more formally as follows. Introduce explicitly the work intensity factor (λ) in the production function. It is a factor that determines the level of output for a given quantity of labor and capital inputs and for a given length of the working period. For the firm j , producing good B , the production function can be written as

$$Q_{bj} = \lambda \Phi_j (D_{hj}, K_{bj}) \quad (4.1)$$

Here D_h is the quantity of labor demanded in the labor market and K_b is the firm's capital endowment.

Assume that the work intensity is endogenous and depends upon the cost that the worker would have to pay if engaged in shirking behavior at the work place. This cost will depend upon the possibility of being caught shirking, which in turn will depend upon the supervision resources used by the firm (assumed to be a fixed cost) and the income loss the worker would suffer if fired from the job. If there were always full employment, that is, if the labor market were Walrasian, the latter cost for workers would be zero because workers who were found shirking and then dismissed from the firm would always be able to find another job at the prevailing market wage rate; hence, workers would have no incentives to supply their highest work effort.

For workers, the cost of shirking is higher if the unemployment rate (u) is higher. But, given the labor market conditions of supply and demand, the rate of unemployment and the real wage rate are not independent of each other; moreover, there is a one-to-one relationship (and a positive one) between unemployment rates and real wage rates.

In order to simplify the model even further, consider only two levels of effort, such as

$$\begin{aligned}\lambda &= \lambda^* \text{ if } u \geq u^* \text{ or } w \geq w^* \\ \lambda &= \lambda' < \lambda^* \text{ if } u < u^* \text{ or } w < w^*\end{aligned}\quad (4.2)$$

Here u^* is the threshold of the unemployment rate above which workers do not shirk (λ^*) and below which they do (λ'). This threshold is socially determined. The degree of partnership in labor relations is assumed to be the determining factor, which is taken as exogenously given. Unemployment is thus a device to get labor discipline. The term w^* is the wage rate that is consistent with the unemployment rate u^* in the labor market. It is also a threshold value leading to labor discipline: the minimum efficiency wage rate.

The individual firm now faces two curves of the marginal productivity of labor, one for each level of effort. They can be represented by the following equations:

$$\begin{aligned}\lambda^* \Phi_{jl}(D_{hj}, K_{bj}) &\text{ if } u \geq u^* \text{ or } w \geq w^* \\ \lambda' \Phi_{jl}(D_{hj}, K_{bj}) &\text{ if } u < u^* \text{ or } w < w^*\end{aligned}\quad (4.3)$$

Summing horizontally over all firms, the aggregate curve of the marginal productivity of labor is determined. This is shown in Figure 4.1(a). The curve $H^*R^*N^*$ represents the marginal productivity of labor when the degree of work intensity is equal to λ^* . The curve $H'R'N'$ represents the marginal productivity of labor when the degree of work intensity is equal to λ' . Because under perfect competition the labor marginal productivity curve also represents the labor demand curve, there seems to be two demand curves, one for each work effort level. Assume labor supply is exogenously given at the value S_h . Therefore, the upper curve requires that the unemployment rate be equal to or greater than the threshold value u^* ; if the unemployment rate is lower than the threshold value, then the relevant curve is the inferior one.

Given the value of u^* , the maximum level of employment (D_h^*) that can maintain the optimal work intensity (λ^*) is determined. This is also shown in Figure 4.1. The segment H^*R^* is the relevant marginal productivity of labor when employment is equal to or less than D_h^* . If employment goes beyond D_h^* , the marginal productivity curve jumps down to the curve $H'R'N'$.

Where is the labor demand curve now? At the wage rate w^* , the quantity of labor demanded would be equal to D_h^* . At levels of real wage above w^* , the quantities of labor demanded are indicated by the segment H^*R^* of the upper marginal productivity curve. At levels of real wage below w^* , would the quantity of labor demanded reach the segment R^*N^* ? No, because this would imply that total employment is beyond D_h^* , which in turn would imply that the unemployment rate would take a value below u^* and shirking would not be as costly as before for workers. Work intensity would fall and the marginal productivity curve would shift downwards to curve $M'N'$, which implies a set of lower profits than at w^* .

Firms would thus have no incentives to pay real wages below w^* . If, for some reason, the market real wage rate were below w^* , firms would have incentives to pay higher nominal wage rates so as to raise the real wage rate and thus to shift the labor

marginal productivity to the upper curve. The real wage w^* operates as a floor price, as a threshold value of the set of efficiency wages. Therefore, of the marginal productivity of labor curve $H^*R^*N^*$, only the segment H^*R^* (the thick line) represents the labor demand curve.

It should be noted that profits could be higher than what they are at point R^* in Figure 4.1(a) if the real wage rate could take a value below w^* but holding constant the total employment level D_h^* . However, this solution would require collective action behavior of capitalists, which is ruled out here due to the selfish rationality. The assumption of the motivation of self-interest does lead capitalists to free-ride behavior, the so called Olsonian problem (Olson, 1965): each firm would seek to hire more workers to make higher profits, but all firms will follow this behavior, and thus more employment would result in the aggregate. Therefore, the vertical segment $R^*D_h^*$ is not part of the labor demand curve. In sum, the segment H^*R^* represents the labor demand curve and along this segment the real wage rate will be determined endogenously.

According to this theory, the labor market cannot operate as a Walrasian market. The solution cannot occur at point N^* in Figure 4.1(a). It appears that firms can make higher profits at point N^* than at point R^* , but this is not the case. At the Walrasian real wage rate, the marginal productivity of labor curve shifts down to $H'R^*N^*$, and the set of profits would be lower than at point R^* . The real wage rate is endogenously determined, but subject to the condition that it must be higher than the Walrasian price.

To be sure, price and quantity of equilibrium in the labor market are still determined by the interaction of supply and demand conditions, but subject to the constraint that the market real wage rate must be equal to or higher than the minimum efficiency wage (w^*). This is equivalent to the condition that the unemployment rate must be equal or higher than a minimum rate (u^*). The minimum efficiency rate of unemployment at the same time determines the maximum employment level that is consistent with the optimum work effort of workers, as judged by capitalists, which may be called *the effective full employment level* (D_h^*). Full employment in the sense of zero rate of unemployment cannot be attained in the labor market.

Assume now that firms have different production functions due to the endowment of specific factors, such as entrepreneurial talent and location. The marginal productivity of labor will now be firm-specific, even if firms are endowed with the same capital stock. Then the curves of marginal productivity of labor may be ordered from the highest to the lowest levels and then aggregated. This is shown in Figure 4.1(b). The aggregated curve is also called H^*N^* , as in Figure 4.1(a), just to show the composition of firms underlying this curve. At the value of real wage w^* , we can see the composition of employment, wage bill, and profits by firms. Profits are the lowest in those firms that are endowed with the lowest specific factors. Hence, the aggregate marginal productivity of labor is falling due to diminishing returns within firms (more workers working with the same stock of capital) and to differential productivity among firms.

In sum, according to this labor market model of the epsilon theory, the labor market in non-Walrasian; labor market equilibrium is necessarily with unemployment. The model assumes that unemployment is the labor discipline device that firms use to maximize profits. Unemployment is a necessity for the functioning of capitalism in epsilon society.

4.3 A Static Model of the Epsilon Theory: The Short Run

In order to make the epsilon theory falsifiable, an epsilon model must be constructed, which can be compared to the task of constructing a particular stage in which social actors will play their roles. The following auxiliary assumptions are introduced for that purpose:

- There are two social groups: capitalists who are endowed with stocks of physical capital and human capital, and workers with stocks of human capital and cash balances. Government behavior consists of supplying money to the economic process.
- The initial nominal wage is given and the rule is that it cannot fall. Wages are paid at the end of the production period; consequently, workers need to hold cash balances for transaction purposes. There is no market for renting capital services.
- Four markets will constitute the market system: labor, commodity, money, and foreign exchange. Human capital is the same for all; then there is only one labor market. The labor market operates with efficiency wages (as shown above). One sole good is produced in society, called good B. Workers and capitalist firms seek to exchange labor services and goods in the labor and commodity markets; money supplied by the government is used as the means of payment. Capitalist firms seek to exchange with international markets, using foreign exchange as the means of payment.
- Capitalists, workers, and the government interact and compete in the market to achieve their goals of self-interest, but none has the power to set prices (a perfect competitive market); they all face costless information on market prices and technology. Cost curves differ by firms due to differences in intrinsic factors, such as entrepreneurship talents and locations.
- The economic process is static and open.

The Behavior of Workers

Workers know that the labor market functions with unemployment: not all workers will be able to exchange the quantities of labor they are willing to sell at the prevailing market wage rate. Once firms have determined the total quantity of labor demanded, the quantity of workers that capitalists will hire from each household is exogenously determined through a random mechanism of exclusion.

The behavior of workers will be similar to what was shown in the Keynesian model in the previous chapter. The reason is that workers will face in epsilon society the same context as in that model. Individual workers are price takers in all markets. As for rationality, it is assumed that the individual workers seek to maximize total

consumption subject to the required real cash balance constraint and the real income constraint.

Just for convenience, the structural equations for worker i are repeated here:

$$\begin{aligned} P_h D_{hi} &= P_b D_{bi} + (D_{mi} - S_{mi}) \\ D_{mi} &= P_b L_i(D_{bi}), L_i > 0, L_i'' < 0 \end{aligned} \quad (4.4)$$

Note that workers cannot decide on the quantity of labor services that they want to sell in the labor market because the labor market is non-Walrasian. Given the values of the exogenous variables, workers will choose the values of the endogenous variables (quantity of good B to buy and quantity of cash balances to hold) according to this rationality. The equilibrium condition is that the cash balances be willingly held ($S_{mi} - D_{mi} = 0$), which implies that total income is spent in buying consumption good B.

The static equilibrium is (trivially) stable. Therefore, the comparative statics method can be applied to derive the effects of changes in the exogenous variables upon the endogenous variables. Thus the individual worker's demand for good B and demand for money can be derived from (4.4); they are

$$\begin{aligned} D_{bi} &= F^i(P_b, D_h, P_h), F_1 < 0, F_2 > 0, F_3 > 0 \\ D_{mi} &= P_b L_i(D_{bi}), L_i' > 0, L_i'' < 0 \end{aligned} \quad (4.5)$$

The Behavior of Firms

The model assumes that the epsilon economy is small in the international economy; so international prices are exogenously given. Epsilon economy exports good B and imports good C, which is a material input required to produce good B. Hence, the domestic price of commodities B and C are equal to

$$\begin{aligned} P_b &= P_e P_b^* \\ P_c &= P_e P_c^* \end{aligned} \quad (4.6)$$

P_e is the nominal exchange rate (the domestic nominal price of foreign exchange) and P_b^* and P_c^* are international prices denominated in units of the foreign exchange. Each domestic price must be equal to the international price multiplied by the nominal exchange rate. If this equality does not hold, people can buy commodities from the cheaper source and sell it wherever is more expensive and make a profit. The ratio $z^* = P_b^*/P_c^*$ is called the *international terms of trade*.

It should be clear that the domestic price level P_b refers to the good produced domestically (good B), which depends upon the exchange rate and the international price of good B. Given the value of the latter, the domestic price level varies according to the variation of the exchange rate. Price stability implies exchange rate stability.

Firms seek profit maximization. Nominal profits is equal to the difference between total revenues minus total costs; total revenues is equal to price times quantity of good B sold in the market; total costs is equal to the sum of the wage bill (nominal

wage times quantity of workers hired) and the total cost of material inputs (domestic price times quantities of good C bought).

Production technology is such that output of good B depends on the services of the stock of physical capital endowed by each firm, the quantity of workers hired, and the material input C bought. Assume that input C enters into the production of good B in fixed quantities per unit of output; that is, assume that input C is not substitutable for capital or labor, although capital and labor are substitutable to each other. Then for every unit of output, the firm must pay a given amount for input C, which is required for technological reasons for the production of good B. Therefore, the *net* value of average labor productivity is now the relevant concept, which is equal to total net value of output (net of the cost of input C) divided by the number of workers; similarly the *net* value of marginal productivity of labor is the relevant concept, which is equal to the additional *net* value of output that results from hiring an additional worker.

The behavior of the individual firm j can then be represented by the following structural equations:

$$\begin{aligned} \text{Maximize} \quad & P_j' = P_b Q_{bj} - P_h D_{hj} - P_c D_{cj} & (4.7) \\ \text{Subject to the production function constraints (where } \lambda^* = 1) \\ & Q_{bj} = \lambda \Phi_j(D_{hj}, K_{bj}), \text{ such that } \lambda = \lambda^* = 1 \\ & \quad = D_{cj} / c \\ & X_{bj} = (1/z^*) D_{cj} \end{aligned}$$

Profits are expressed in nominal terms. Total output is net of depreciation, so is total profits. The term c is a technological coefficient, equal to the number of units of good C required to produce one unit of good B. Imports are all intermediate goods. The production function is subject to *limitational factors* and is thus represented as a system of equations.³ The last equation answers the question: who does the firm pay for the imported input C? Because good C is an imported input, the firm pays foreign firms with foreign exchange that obtains by exporting part of its output B (X_b) according to the terms of international trade z^* .

The structural equations presented above lead to the following equilibrium condition:

$$\begin{aligned} \Phi_{j1}(D_{hj}, K_{bj}) (P_b - c P_c) &= P_h, \text{ or} & (4.8) \\ \Phi_{j1}(D_{hj}, K_{bj}) (1 - c/z^*) &= w, \text{ where } (c/z^*) < 1 \end{aligned}$$

The marginal productivity of labor is now net of the cost of using the limitational input C. Profit maximization implies the equality between the value of the net marginal productivity of labor and the nominal wage rate or, alternatively, the equality between the net physical marginal productivity of labor and the real wage rate. Profits arise from

³ A *limitational factor* is defined as follows: an increase in its quantity is a necessary but not a sufficient condition for an increase in the level of output. This definition implies that this factor of production cannot be substituted by the others. In the technology assumed here, labor and capital can be substituted one for the other, but there is no substitution between input C and the combined factor labor-capital.

the difference between the average productivity of labor and the marginal productivity of labor multiplied by the employment level.⁴

Individual firms are price takers in all markets. Therefore, for each firm, profit maximization implies hiring workers up to the point that the value of the net marginal productivity of labor is equal to the market nominal wage rate. Given the exogenous variables (the nominal market prices and the capital stock), firms will choose to produce a quantity of good B, buying a quantity of input C and hiring a quantity of workers. As long as the exogenous variables remain fixed, these quantities of the endogenous variables will be repeated period after period. This is the static equilibrium in the behavior of a typical firm.

The equilibrium situation is (trivially) stable. Comparative statics can then be applied to determine beta propositions about the behavior of the firm. The critical variables are the quantity of labor demanded and the quantity of good B supplied to the domestic market. They are

$$\begin{aligned} D_{hj} &= H^j(P_b, P_c, P_h, K_b), H^j_1 > 0, H^j_2 < 0, H^j_3 < 0, H^j_4 > 0 \\ \text{Subject to} \\ P_b S_{bj} &= P_b Q_{bj} - P_b X_{bj} - P'_j = (P_c D_{cj} - P_b X_{bj}) + P_h D_{hj} \end{aligned} \quad (4.9)$$

At the firm level, the quantity of labor demanded depends positively on both the price level and the capital stock because an increase in the value of these variables increases the value of the marginal productivity of labor; it also depends negatively on the price of the input C because an increase in this price reduces the net value of the marginal productivity of labor; finally, it depends negatively on the nominal wage rate because an increase in this price would lead the firm to increase also the value of the marginal productivity of labor, which in turn would imply a fall in employment (due to diminishing returns). The second equation is the budget constraint of the firm, which shows that what the firm sells in the market must allow the firm to pay for the purchase of the production factors. The quantity supplied of good B to the domestic market is derived from this equation.

The Behavior of the Government

In this model, politicians will constitute a social class. As any other individual, they will act guided by the motivation of self-interest. Politicians will not be considered special people, whose actions are motivated by altruism. In democratic systems or even in some forms of authoritarian system, the legitimacy of political power comes from public support in the form of votes or public opinion. Therefore, politicians will seek a place in government by investing in the electoral competition; once in government, they will seek to maintain political power and incomes. They will give priority to the objective of

⁴ Dimensionally, the ratio c/z^* is a pure number. The condition that $(c/z^*) < 1$ assures that the technological system is productive. If a firm exchanges one unit of good B for 2 units of good C as inputs, which help to produce less than one unit of good B, the only thing the firm has done is to run down its stock of good B. For the system to be productive these 2 units of good C should help to produce, say, four units of good B. Then the firm starts with one unit of good B and ends up with four units of it.

maintaining their position in the political class. The theory of government behavior is thus much alike the theory of capitalist behavior.

This rationality also implies a hierarchy on government policies. Policies that have effects in the short run (and will influence the voting in the next elections) have priority over those that have long run effects (when the influence on the next elections is not significant). Government behavior will then be myopic and short sighted to address the long run problems of society.

A simple model of the government theory can be constructed using auxiliary assumptions. Assume that governments seek to maximize votes, subject to several restrictions: the public budget, the force of pressure groups upon the public budget, and the mandatory expenses associated to the financing of rights. On the government budget, the assumption is to consider money supply as the only source of government income. The government prints money and transfers it to individuals in the lump sum form, as if were by helicopters. This mechanism is known as “helicopter money.” Other income sources such as taxes and debt will be ignored in this model.

Now suppose that voting for the current government depends positively upon the indicators of the economic process. Higher national income level (which implies higher employment level and lower unemployment level), higher real wages (or lower inflation rate) will increase the voting support to the government. The only instrument the government has is the supply of money. The government will thus respond to the outcomes of the economic process with changes in the quantity of money according to its own interest. For example, if there is a recession, and output and employment fall, the government will seek to stimulate the economy and supply more money. If for any reason, inflation is too high, the government will change the supply of money to reduce it.

The political cycle will also change the government expenditure. The government will seek to buy votes by increasing the money supply during electoral periods.

In sum, government behavior is also guided by the motivation of self-interest; therefore, the supply of money will be endogenous. The government is not above the class relations, but rather, interacts with the social classes following its own interests. Therefore, its social function is accomplished as a by-product of its selfish motivation. This theory was firstly proposed by Downs (1957) long time ago. (A more elaborated model of government behavior will be developed in Chapter 7 below.)

4.4 Market General Equilibrium

It is now time to aggregate the individual behavior of social actors. Aggregating over all workers, the market demand functions for good B and for money are determined. This is simply the result of adding all individual functions, for workers make their choices independently, not by group interactions or collective actions.

These functions are presented as the following structural equations:

$$\begin{aligned}
D_b &= F(P_b, D_h, P_h), & F_1 < 0, F_2 > 0, F_3 > 0 \\
D_m &= P_b L(W), & L' > 0, L'' < 0 \\
\text{Subject to the aggregate budget constraint of workers} \\
P_h D_h &= P_b D_b + (D_m - S_m)
\end{aligned} \tag{4.10}$$

The quantity of good B demanded depends negatively on the price level, positively on the nominal wage rate, and positively on the level of employment of the economy; that is, it depends on the quantity of workers that firms are willing to employ, not on the quantity of labor that workers are willing to sell. The higher the employment level, the higher the quantity of good B demanded will be. (In a Walrasian market, where workers sell all their labor supply, their total income depends on nominal wages only.) The demand for money depends on the price level and the wage bill because the demand comes from workers alone.

Aggregating over all firms, the market behavior of the firms can be written as the following set of structural equations:

$$\begin{aligned}
D_h &= H(P_b, P_c, P_h, K_b), & H_1 > 0, H_2 < 0, H_3 < 0, H_4 > 0 \\
\text{Subject to} \\
P_b S_b &= (P_c D_c - P_b X_b) + P_h D_h \\
X_b &= (1/z^*) D_c = (1/z^*) f(D_h, K_b), & f_i > 0
\end{aligned} \tag{4.11}$$

The first equation is the labor demand function. The second equation shows the quantity of good B supplied to the domestic market, which is derived from the firms' budget equations. The third equation represents the condition that firms must pay for the import of input C to foreign firms, that is, they must export good B in the amount X_b . This is the foreign exchange market equilibrium condition.

From the structural equations shown above, the general equilibrium conditions can now be established. They are

$$\begin{aligned}
\text{Labor market} & \quad S_h \equiv D_h + U, & P_h \geq P_h^*, w \geq w^* \\
\text{Good B market} & \quad S_b = D_b \\
\text{Money market} & \quad S_m = D_m \\
\text{Foreign exchange} & \quad P_b^* X_b = P_c^* D_c \\
\text{Subject to the aggregate budget constraint} \\
P_b (D_b - S_b) &+ P_c (P_c^* D_c - P_b^* X_b) + (D_m - S_m) = 0
\end{aligned} \tag{4.12}$$

From the four markets, one is non Walrasian (labor market) and the rest are Walrasian. The equilibrium condition in the good B market refers to the domestic market, in which the demand comes from workers only; the equilibrium condition in the money market states that the quantity of money supplied by the government must be willingly held by workers; and the equilibrium condition of the foreign exchange market comes from the budget equation of firms.

From the aggregate budget constraint of system (4.12), Walras' law is derived: one of the three Walrasian market is redundant. Thus the system needs to solve only for three markets: two Walrasian and the non-Walrasian.

The general equilibrium must solve for the endogenous variables of the system for given values of the exogenous variables. These variables can be separated as

Endogenous variables: $P_e, D_h, P_b, P_c, D_b, S_b, D_c, X_b, U, Y, D$

Exogenous variables: $z^*, r^*, K_b, S_h, \delta, S_m$

Total output (Y) and its distribution (D) are endogenous. The initial degree of inequality δ appears explicitly in the set of exogenous variables. The international rate of interest r^* has been added to the set of exogenous variables, the justification of which will be provided later on, when the banking industry is introduced. Money supply (S_m) appears as exogenous, but it belongs to the special category of quasi-exogenous, as will be discussed later on. Notice that some of the endogenous variables at the general equilibrium level of analysis were exogenous at the microeconomic level, such as the nominal prices. This is in accord with the principle of increasing endogenization of variables in the aggregation process.

In this general equilibrium model, the output and distribution of equilibrium is the result of the interactions between the three social actors of the model: capitalists, workers, and the government. Just for the sake of simplicity, the general equilibrium solution will be found sequentially: given the quantity of money, the market system determines total output and its distribution; then the government acts endogenously to modify the market result by changing money supply so as to achieve another general equilibrium with higher output level and higher real wage rate, subject to the constraint of having low inflation rates and the demands of the pressure groups.

The task is now to identify the core of the general equilibrium system. Start with foreign exchange market. Given that this is a one-good economy, in which imports are production inputs, the condition of equilibrium is very simple: firms must export good B to pay to foreign firms for the imported input C. The higher the exchange rate, the higher the price level, and the higher the quantity of labor demanded and also the quantity of input C, which implies a higher quantity of foreign exchange demanded; therefore, the demand curve for foreign exchange is upward sloping. But the quantity supplied of foreign exchange will always be equal to the quantity demanded because firms import and export at the same time and thus decide for the quantity demanded and supplied of foreign exchange. The nominal exchange rate will be determined in the money market.⁵

⁵ Under this system of flexible exchange rates, a particular rule for the functioning of the foreign exchange market can be assumed. Firms export commodity B and get foreign exchange, which they sell to the central bank and get domestic currency; then firms use domestic money to buy back foreign currency, which they use to import commodity C. Transaction costs in these operations are zero. In this oversimplified model, the same firms are the actors behind demand and supply in the foreign exchange market. This market looks artificial because there is only one domestic good. In a multi-good economy, in which imports include production inputs and final goods, the agents that import will be different from those that export and the equilibrium in the foreign exchange market will require to solve a coordination problem, through markets, because demand and supply conditions are determined independently. But the conditions of equilibrium will be similar. Problems of foreign debt are ignored in this model.

From the labor demand function, it follows that the labor demand curve has the property that at each nominal wage rate there will exist a quantity of labor demanded, which also determines the quantity of input C and therefore the quantity of foreign exchange demanded. Since the same firms must pay for input C by exporting part of output B, the quantity of foreign exchange demanded must be equal to the quantity supplied. Therefore, equilibrium in the foreign exchange market will exist at each employment level. The equilibrium condition in the foreign exchange market is then already incorporated into the labor market. The assumptions of input C as limitational factor and one-good economy allow us to reach this conclusion.

The core of the general equilibrium system is then constituted by two markets alone: the labor and money markets. The core system can be written as follows:

$$\begin{aligned} D_h &= H(P_e; P_h, z^*, K_b), H_1 > 0, H_2 < 0, H_3 > 0, H_4 > 0 \\ S_m &= P_e P_b^* L(W) = M(P_e, D_h; P_h, P_b^*), \text{ where } M_i > 0, \text{ for all } i \\ \text{Subject to} \\ D_h &< D_h^* < S_h \end{aligned} \quad (4.13)$$

The first equation is the labor demand function; the second is the equilibrium condition in the money market. These two equations solve for the two endogenous variables, the employment level and the nominal exchange rate. The constraints indicate that we are searching for the existence of general equilibrium with unemployment.

The core of the general equilibrium is shown in Figure 4.2, panels (a) and (b). The labor market in terms of the nominal wage rate is presented in panel (a). The value of the gross marginal productivity of labor is represented by the curve $h'n'$ and the value of the net marginal productivity of labor by the curve hn . The gap between these two curves measures the cost of using the limitational factor C, which is paid with exports of good B. Given the nominal wage rate, the labor market will determine the employment level, which must not exceed the level D_h^* .

The money market equilibrium can be established using the exchange rate instead of the price level, as shown in panel (b). The higher the exchange rate, the higher the domestic price level, and, thus, the higher the demand for nominal money. A downward sloping demand curve for money can be obtained by using $1/P_e$ as the "price of domestic money in units of the foreign exchange". Given an initial stock of money, there will be an exchange rate value that clears the money market.

It can be seen graphically that the solution of the core is simultaneous. The labor market must solve for the employment level, which requires knowing the foreign exchange value. The money market must solve for the foreign exchange rate, which requires knowing the employment level. There is one, and only one, set of values of the exchange rate and the employment level that can produce equilibrium in both markets. Point e and point F show the core solution in Figure 4.2.

From the solution of the core, the equilibrium values of the rest of the endogenous variables can be found by implications only. So once the exchange rate is known, the price level is also determined. Therefore other real variables, such as the real wage rate, are also determined.

The labor demand curve related to the real wage rate is presented in panel (c). The curve H'N' represents the gross marginal productivity of labor and the curve HN the net marginal productivity of labor. The gap between these two curves measures the quantity of exports of good B needed to pay for the use of the required quantity of imported input C. Therefore, at each employment level there is equilibrium in the foreign exchange market; moreover, the labor demand curve now has another parameter: the international terms of trade z^* .

Because the work intensity level depends on the unemployment rate, panel (c) also shows that there is a minimum unemployment rate (u^*) that assures this intensity. Given the labor supply, u^* implies a minimum wage (w^*) for the set of efficiency wages and also the maximum employment level (D_h^*); therefore, only the segment HR (of the curve HN) represents the labor demand curve.

The real wage rate is consistent with the employment level of equilibrium found in the core. In panel (c), equilibrium occurs at point E. The area under the curve H'E' measures total output and the area under the curve HE total output net of exports. The gap measures the quantity of good B exported to pay for the required quantity of input C in producing the gross output. Thus the foreign exchange market is also in equilibrium.

The area under the curve HE (not under H'E') measures total net output, net of depreciation and net of the cost of imported inputs. This is distributed among workers (the wage bill W) and capitalists (profits P). This is the income distribution of equilibrium. The commodity market should also be in equilibrium due to Walras' law, which can be confirmed graphically: the area showing total quantity of commodity B supplied to the domestic market (equal to total net output minus profits retained by capitalists) is equal to the area of the total wage bill, which is equal to the quantity demanded for commodity B domestically. The market demand for good B comes only from wage earners.

Production and distribution of equilibrium have thus been solved. The total output of equilibrium is called the national income (Y). The distribution of this income between social classes, as total profits and total wages, can be written as follows:

$$\begin{aligned} Y &= P + W = P + w D_h \\ S_h &= D_h + U \end{aligned} \tag{4.14}$$

In order to make it clear that the capitalist system operates with unemployment, the labor allocation equation includes the size of unemployment (U) with zero wage income.

To have social viability in the general equilibrium, it is assumed that the unemployed get incomes from the public budget as unemployment insurance, part of economic rights established in epsilon society. To maintain the same efficiency level, the transfer per worker will have to be smaller than the market real wage rate.

Two concepts of income distribution need to be distinguished now: The *functional distribution*, which refers to the shares of profits and wages in national

income; and the *personal distribution*, which refers to the distribution of individual incomes. The latter takes into account the transfers of income to the unemployed through the fiscal budget. The functional distribution will not change with the fiscal policy if the same tax rate is applied to the wage bill and to total profits. Assume that unemployment insurance is set as a given fraction of the market wage rate; then the inequality between the employed and the unemployed will not change.

The general equilibrium solution is an equilibrium situation because no one has both the power and the incentive to change it. The values of the endogenous variables will be repeated period after period as long as the values of the exogenous variables remain fixed.

Equilibrium output and distribution will be the result of the interactions between the three social actors of the model: capitalists, workers, and the government. A fourth social actor is the foreign producers of good C, but its behavior affects the domestic economy through changes in the international terms of trade.

As stated before, this is a theory about the market system behavior, which assumes that the market system is able to solve this problem. Obviously the theory does not say that markets solve for prices and quantities by actually solving equations; it rather says that whatever mechanism markets have to solve this problem (auctioneers, telecommunications, brokers, trial and error, etc.) is equivalent to solving equations. The theory says that market system operates *as if* it solved a system of equations, as if it were a big computer.

Macroeconomics textbooks consider models in which total unemployment is analytically decomposed into excess labor supply and frictional unemployment (Krugman and Wells 2006). Frictional unemployment is the result of job matching problems: because workers and jobs differ, workers search for jobs and firms search for workers, at the Walrasian wage rate in the labor market. In this epsilon model, by comparison, frictional unemployment will be ignored because all workers and all jobs are alike and job search problems could hardly exist or will be very small; thus even in that situation, general equilibrium is with unemployment, which is equal to excess labor supply. Unemployment as excess labor supply is thus a structural characteristic, a necessity, for the functioning of the epsilon society.

The assumption of a minimum unemployment rate made in the labor market model of the epsilon theory has also been introduced in some macroeconomics textbooks, from which the “natural rate of unemployment”—now defined as the non-increasing inflation rate of unemployment—has been derived (Blanchard 2009, Chapters 6 and 8). The necessary unemployment rate (u^*) of the epsilon model is thus equivalent to the natural rate of unemployment in the sense that in both cases an increase in money supply (or in any other nominal variable) will increase the rate of inflation without reducing the rate of unemployment.

4.5 Empirical Predictions

It is relatively easy to show that the core of the general equilibrium is stable. In the labor market (a non-Walrasian market) stability is trivial, as firms will be able to

readjust automatically if the quantity of employment is for some reason out of equilibrium; that is, there cannot exist instability in the labor market. Therefore, the only place where instability would possibly exist is the money market (a Walrasian market). The money market is indeed stable because the demand curve is sloping downwards and the supply curve is a vertical line; hence, any price that is for some reason out of equilibrium would restore its equilibrium value automatically. Stability of the general equilibrium requires stability in the money market only, which is fulfilled. Therefore, comparative statics can be applied to the general equilibrium solution to derive beta propositions.

The effect of changes in the exogenous variables upon the endogenous variables in the short run can now be analyzed. The relevant exogenous variable include only the international terms of trade (z^*). Remember that this model assumes that government monetary policy is endogenous, a reaction to external shocks.

Two equilibrium situations must then be analyzed, one with unemployment rate above the necessary rate ($u > u^*$) and the other equal to it ($u = u^*$). Full employment is usually defined as excess labor supply equal to zero ($D_h = S_h$). This is unattainable in epsilon society. What is relevant is the *effective full employment* ($D_h^* = (1 - u^*)S_h$), which is necessarily smaller than the quantity supplied of workers.

An increase in the international terms of trade (z^*) shifts the labor demand curve outwards, as can be deduced from Figure 4.2 (c). Considering that the initial equilibrium is at point E, firms would now seek to hire more workers; hence, the wage bill would increase, which would imply a higher quantity of real cash balances demanded; but given the quantity of money supplied, the build-up of cash balances would imply a temporal fall in the quantity of good B demanded, which would increase both exports and the inflow of foreign exchange, which would lead to a fall in the exchange rate, which in turn would lead to a fall in the domestic price level and a rise in the real wage rate; hence, employment would tend to fall, but not to offset the initial increase. As a result of these adjustments, both employment and real wage rate will increase. Thus the initial equilibrium situation, given by point E, will be moved to another equilibrium situation, say point E'' (not marked), which will be located to the north-east of point E.

The effect of an increase in the international terms of trade upon income inequality can also be seen in Figure 4.2 (c). The new equilibrium implies higher output level, higher wage bill, and higher profits; hence, the change in income inequality will be ambiguous.

The epsilon model assumes that the supply of money is "endogenous," but in a particular sense: the government reacts to the results of the market system, seeking to alter these results in the direction indicated by the logic of maximization of votes. Money supply is the only instrument that the government can manage to achieve this objective. Money supply is not purely exogenous or endogenous; hence it could be classified as *quasi-exogenous* variable in this model.

An increase in the money supply generates an excess of cash balances willingly held by workers, who will then intend to restore the equilibrium by using this excess to buy more quantities of good B in the period of adjustment. This adjustment would reduce both the exports and foreign exchange inflows and thus an increase in the

exchange rate. Then the price level would increase and consequently the real wage would fall; hence, firms would seek to hire more workers and produce more output to gain more profits. This would be the first round effect. The second round effect refers to the feed-back effects. A fall in the real wage rate and an increase in the employment level imply an ambiguous change upon the real wage bill and thus upon the demand for real cash balances and upon the price level. Therefore, the second round effect upon the price level could be small and may be neglected. In sum, the effect of an increase in money supply is to decrease the real wage rate and to increase the employment level.

The effect of an expansionary monetary policy can be seen in Figure 4.2(c). Point E represents the initial general equilibrium situation. An increase in the money supply increases the price level, which implies a fall in the real wage rate. The employment level increases and total output also goes up. The initial equilibrium situation, at point E, will be moved to another equilibrium situation, which will lie below point E, along curve HN. The government can then reduce unemployment using monetary policy.

A sufficient large increase in money supply will ultimately reach the effective full employment level. The equilibrium situation can move from point E to point R. If money supply keeps increasing, employment will remain fixed, and only the exchange rate and consequently the price level will rise, which will then induce increases in the nominal wage rate in the same proportion in order to avoid the fall in the minimum efficiency wage rate (w^*); that is, money becomes neutral at the effective full employment level. It should be clear that the Walrasian market equilibrium, point N, is unattainable in the epsilon model.

Government behavior implies that the *net effect* of the exogenous variable international terms of trade upon endogenous variables include the direct effect and the induced effect upon money supply. Consider the following equations:

$$\begin{aligned} Y &= F(z^*, S_m), F_1 > 0 \\ S_m &= H(z^*), H' < 0 \\ Y &= F(z^*, H(z^*)) = F_1 + F_2 H' > 0 \end{aligned} \tag{4.15}$$

The first equation indicates that, in the short run, the level of output Y depends positively upon the terms of international trade and the quantity of money. The second equation says that government behavior responds to this external shock by changing the money supply. Vote maximization behavior would lead the government to have counter-cyclical monetary policies: an external negative shock causes economic recession, which induces the government to increase money supply to counter act this shock. The third equation assumes that the external shock can be offset only partially with monetary policy; hence, the model assumes that the direct effect (the sign) of the terms of international trade will prevail.

In reality, the model of vote maximization behavior predicts that government policy would seek to maintain effective full employment, that is, limit unemployment to only the necessary unemployment rate (or the “natural rate of unemployment”). Higher rates of unemployment would be the result of other shocks, such as foreign debt crisis, which is ignored in the epsilon model presented here.

In the epsilon model, the government has two policy instruments: money supply and the nominal exchange rate, but only one degree of freedom: once one of them is chosen, the other is endogenously determined. If the government uses money supply as a policy instrument, the exchange rate is determined endogenously (as chosen in this model). But the government could use the exchange rate as instrument, fix it, and then the money supply will be determined endogenously, according to the needs of cash balances of the social actors, as indicated in the second equation of system (4.13).

The government can have other instruments, depending on the model. If we introduced the bond market in the model, the interest rate (r) would be another endogenous variable. Given that in an open economy there would be free mobility of financial capital, the international interest rate (r^*) would appear as another exogenous variable. Foreign exchange will move in or move out of the economy depending, among other things, upon the differentials between domestic and international interest rates. The decrease in the international interest rate increases will have the same effect of an increase in the international terms of trade: the inflow of foreign exchange will rise; then both the exchange rate and the price level will tend to fall.

In this case, the government would have three possible policy instruments: the quantity of money, the nominal exchange rate, and the interest rate. However, there would still be one degree of freedom only: the government can choose one of them, and the other two will be determined endogenously.

In the epsilon model, government behavior can affect total output and distribution using monetary policy only ("helicopter money" only). However, the model is not as restrictive as it appears. In other models the government could be using other nominal variables as instruments, but the relation between nominal variables and real variables will correspond to that predicted by the epsilon model: in the short run changes in nominal variables affect the real variables, but not up to the situation in which equilibrium with full employment is reached. General equilibrium is with positive rate of unemployment in the epsilon society.

A new exogenous variable appears in the general equilibrium model as a consequence of aggregation: the initial inequality (δ). Its effects are analyzed now.

Changes in the initial inequality would occur if the distribution of individual endowments of economic assets is redistributed exogenously. An increase in this variable implies a higher degree of concentration of economic assets. This would occur if the capitalist class becomes smaller in size or, given its size, the concentration of the capital stock increases within the class. The capitalists that loose capital stock ownership will become workers and continue receiving wage income, as they are endowed with human capital.

The effect of an increase in the initial inequality upon output will be nil because production efficiency is independent of the degree of inequality in society (an assumption to be relaxed later on). As a result, the functional distribution of total output will not change either; however, the personal distribution of income will increase because the same amount of profits will go to fewer capitalists. In all the cases analyzed before, functional and personal distribution move in the same direction, except in this case. The relation between changes in the initial inequality in assets and the degree of inequality in the distribution of the income flow will be direct: the higher the inequality

in the distribution of the *stock* of capital, the higher the inequality in the personal distribution of the *flow* of incomes.

The reduce form equations for total net output or national income (Y) and the degree of inequality in its distribution (D) derived by the comparative statics method are as follows:

$$Y^0 = F^{\varepsilon}(z^*, r^*, \delta; K_b, S_h), \quad (4.16)$$

$$F_1 > 0, F_2 < 0, F_3 = 0, F_4 > 0, \text{ and } F_6 = (?)$$

$$D^0 = G^{\varepsilon}(z^*, r^*, \delta; K_b, S_h), \quad (4.17)$$

$$G_3 > 0, \text{ and } G_i = (?) \text{ otherwise}$$

Equation (4.16) indicates that factor endowments determine the level of national income in the long run, with short run variations due to the effect of changes in two exogenous variables: the international terms of trade and the international interest rate. Similarly, equation (4.17) indicates that the initial inequality determines the level of inequality in the distribution of national income, with short run variations around that level due to changes in the other exogenous variables.

It should be noted that, from the way the epsilon model has been constructed, the macro behavior of the epsilon economy shown above has micro foundations and that, at the same time, the underlying micro behavior of social actors has macro foundations. The model thus provides unity of knowledge. The partial derivatives of the system (4.16)-(4.17) constitute the causality relations or the beta proposition or the empirical predictions of the epsilon model. (The mathematical derivation is presented in Appendix A.) They can be used in the process of falsification of the epsilon theory.

4.6 Empirical Consistency: The First World Countries

Does the abstract epsilon society resemble well the group of First World countries? The most notable empirical regularities of the First World countries in the short run were listed in Chapter 2. The beta propositions of the epsilon static model presented here can now be confronted against the relevant facts, which include Facts 1 and 4.

The epsilon model predicts that any equilibrium situation will imply a positive rate of unemployment ($u > 0$) because unemployment plays a role in the economic process: capitalism needs unemployment to operate. This equilibrium condition is observable and, thus, it makes the model refutable. Thus the epsilon model indeed predicts the existence and persistence of unemployment (Fact 1). The epsilon model also predicts the interplay between nominal and real variables in the short run (Fact 4). According to the epsilon model, changes in the money supply (or in the nominal exchange or in the nominal interest rate) will affect output and employment.

The predictions of the epsilon model are also consistent with the empirical results of modern macroeconomics. This literature explains the existence of positive unemployment rate by using also the assumption of efficiency wages in the labor market. In this literature, the modified Phillips curve shows an inverse relation between changes in the inflation rate and unemployment rate, which in the case of the U.S.

economy, for the period 1976-2008, shows that the *minimum* unemployment rate observed is 4%. This relation and the non-zero minimum unemployment rate would also be predicted by the epsilon model if the inflation rate were utilized instead of the price level. Moreover, the *average* unemployment rate for a long period (1970-2008), varies across countries: 2.3% in Japan and 6.1% in the United States (Blanchard 2009, p. 190). The epsilon model would explain these figures by the differences in the necessary unemployment rates (u^*) between these countries: labor relations are less conflictive in Japan than in the US.

Vote maximization rationality of governments predicts that governments will seek to apply policies to maintain effective full employment, which is equivalent to maintain unemployment at the necessary rate only. Facts indicate a certain degree of consistency with this prediction. The observed annual rates of unemployment in most capitalist countries show variations within relatively narrow ranges in the period 1960-2005: between 3% and 10% in the US and 2% and 11% in Western Europe (shown as Fact 1 in Chapter 2). Short run factors other than changes in international prices (terms of trade and interest rates) have been ignored in the epsilon model and may be the cause of those fluctuations above the necessary unemployment rate.

According to the epsilon model, unemployment in the labor market is a necessity, not a possibility, as predicted by the neoclassical, Keynesian, and classical theories. On the other hand, labor markets do not operate with subsistence real wages, as the classical theory assumed. If real wages covered the needs of workers, poverty would hardly exist in the First World. According to the epsilon model, workers do not earn what they need for living; rather they make their living with what they earn. Limits to the real wage rate are determined by efficiency wages, not by subsistence wages.

In sum, the two fundamental facts of the First World countries (Facts 1 and 4) do not refute this model of the epsilon theory; hence, there is no reason to reject epsilon theory and we may accept it provisionally at this stage of our research. Theories that intend to explain production and distribution in the Third World will be presented in the next two chapters.

Figure 4.1. Labor Productivity and Labor Demand

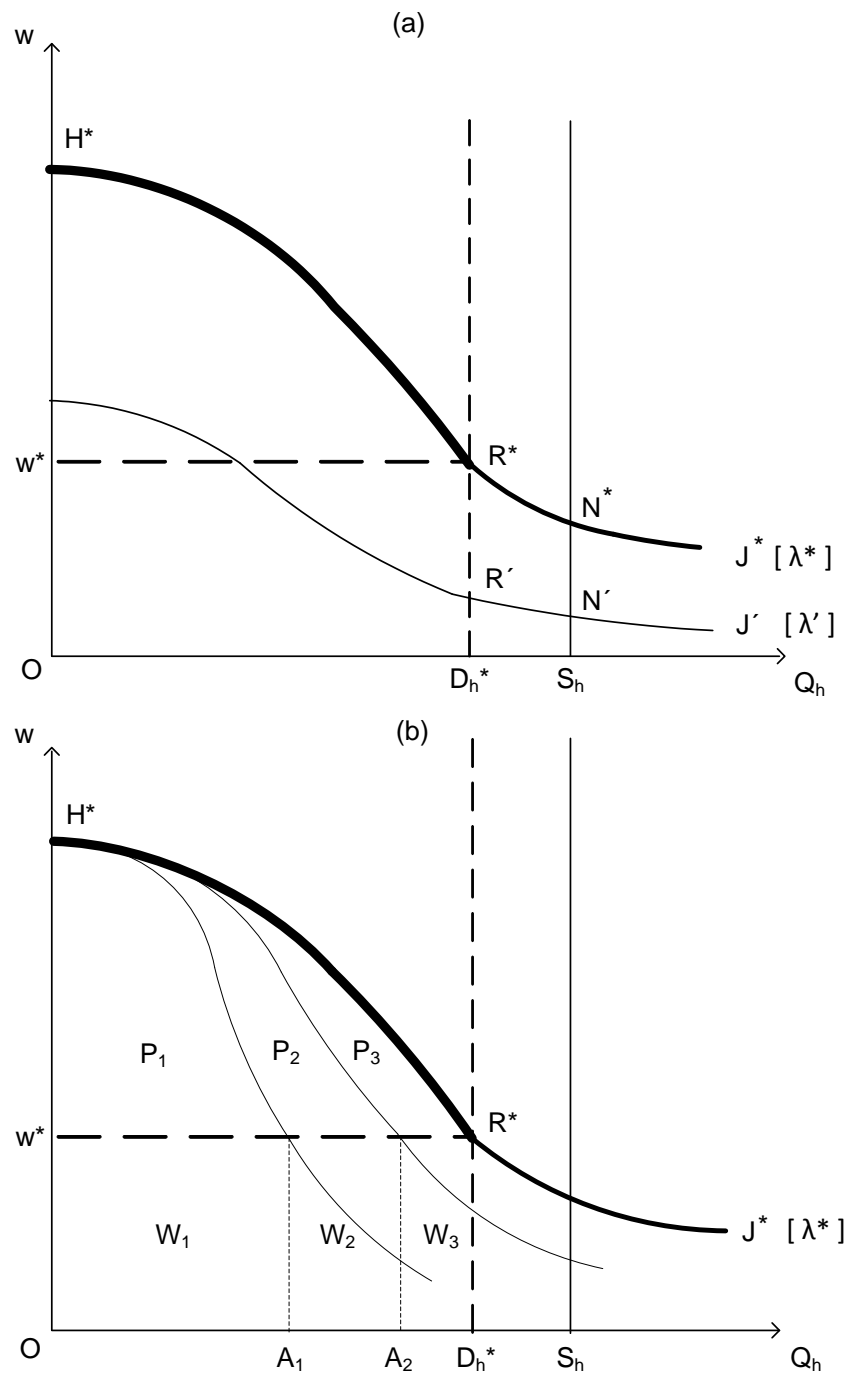
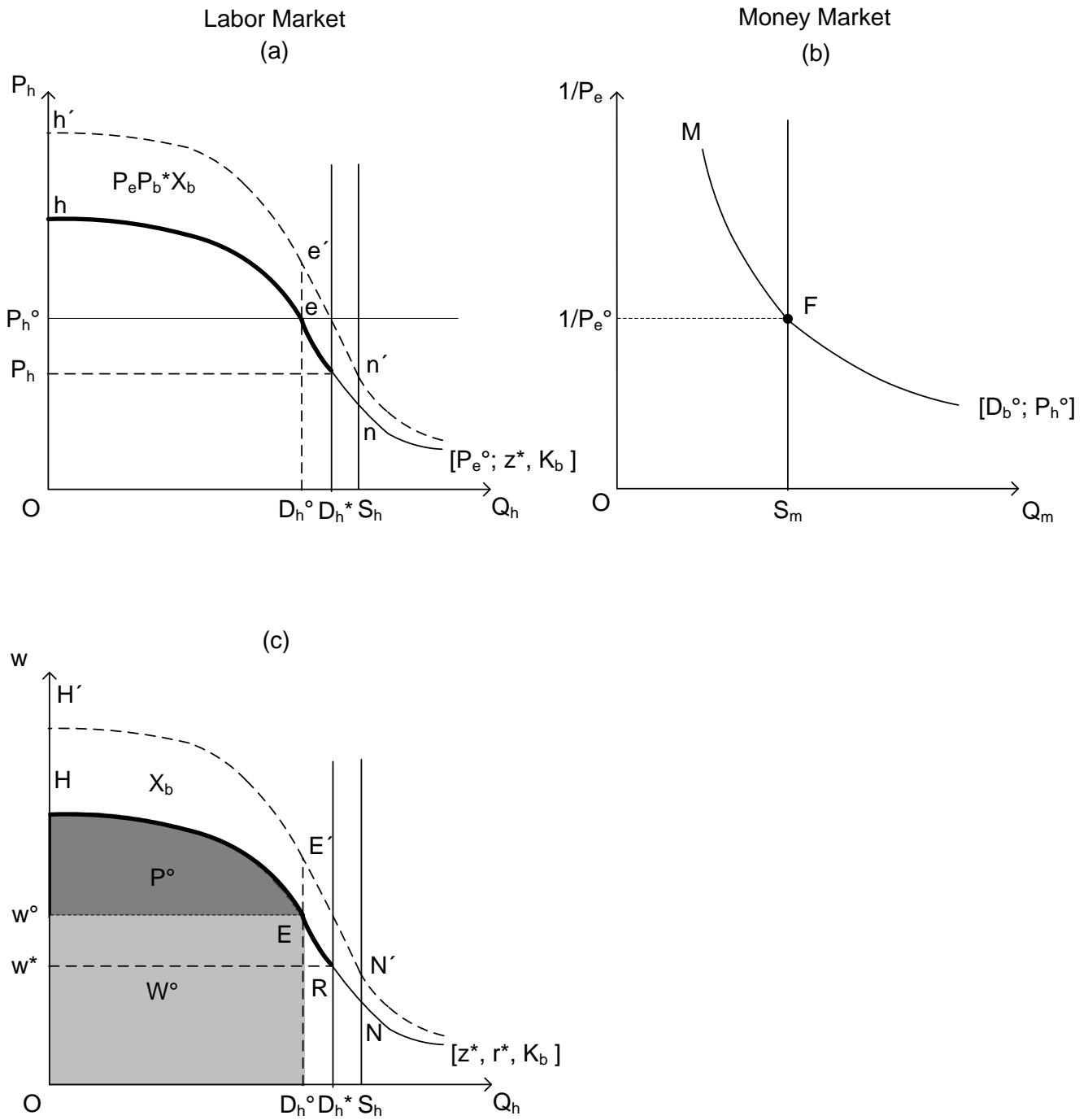


Figure 4.2. General Equilibrium in the Open Epsilon Society



CHAPTER 5

THE THIRD WORLD

In Third World countries, in addition to unemployment, there exists underemployment in the self-employment sector. Workers who are out of wage employment are mostly engaged in self-employment. They produce goods and services in small units of production that are located in rural areas (in small farms), and in urban areas (in small shops, some of them on the streets). The income they make in these small units is, in general, lower than the average wage rate being paid in the labor market for similar skills. This is Fact 2 listed in Chapter 2.

How can one explain production and distribution in Third World countries? Surely, by constructing a theory, an abstract society. An abstract capitalist society—called omega—will be constructed now and then submitted to empirical refutation.

5.1 Omega: Socially Homogeneous Class Society and Overpopulated

Omega society will also constitute a class society. Its social institutions are thus similar to those of the epsilon society. The only difference rests upon factor endowments: omega is an overpopulated society.

Given its technological knowledge, the labor productivity level in a society depends upon its factor endowments. The higher the stock of capital per worker in society, the higher the average labor productivity (output per worker) will be. Assume that omega is endowed with lower capital per worker than epsilon. This could be the result that labor supply increases exogenously and significantly in the epsilon society. As the size of the labor supply increases, capital per worker will fall and so will the average labor productivity, which entails also a fall in the marginal labor productivity. Consider that population has increased in such magnitude that the marginal productivity of the total labor force is zero, which is the definition of an *overpopulated society*, and we have an omega society. The implication of overpopulation is that if a portion of the labor force could be displaced from the production process, total output would not change. By way of comparison, epsilon society was defined as an under-populated society, in which the marginal productivity of the total labor force is largely positive.

Could omega society operate as an epsilon society? As shown in the previous chapter, epsilon operates necessarily with unemployment; hence the only effect of an increase in the labor supply would be the increase in unemployment. If the labor supply continued to increase, unemployment would be so large that the labor market would not be

able to function with the prevailing institutional norms. If the government were paying unemployment insurance, the financing of this program would become unviable. The excess labor supply would be well beyond the necessary rate of unemployment (u^*); so unemployment would be too large for the unemployed workers to make a living just by searching for jobs and finding jobs.

In sum, in an overpopulated economy, unemployment cannot operate as the device to discipline workers. Therefore, omega will function differently from epsilon. To show how the omega society will function is the objective of this chapter.

The alpha propositions for the omega society (ω) include all the propositions that were established for the epsilon economy, except that, as initial conditions, there will exist overpopulation in the labor market. For easy reference the set of alpha propositions are fully listed as follows:

$\alpha(\omega).(1)$ *Institutional context*: (a) Rules: People participating in the economic process are endowed with economic and political assets. Economic assets are subject to private property rights. Individuals can exchange goods subject to the social norms of market exchange, which include the norm that nominal wages in labor markets cannot fall. Individuals are entitled with equal political rights and duties; hence, there is one single class of citizenship in the democratic system. (b) Organizations include households, firms, and the government.

$\alpha(\omega).(2)$ *Initial conditions*: (a) Individuals are endowed with unequal quantities of economic assets but with equal political entitlements. There are two social classes: capitalists and workers. (b) Factor endowments: the stock of capital per worker is such that the marginal productivity of the entire labor force is zero, that is, there is overpopulation. There exist two production sectors: the capitalist and the self-employment sector.

$\alpha(\omega).(3)$ *Economic rationality of agents*: Individuals act guided by the motivation of self-interest. Capitalists seek two particular objectives: maintenance of class position and profit maximization, such that the former has priority. In the labor market, workers seek to maximize wages and minimize effort; hence, capitalists use devices to extract effort from workers. Due to this conflict in labor relations, capitalists use particular devices to extract effort from workers, which cannot be based on unemployment.

5.2 A Static Model of the Omega Theory: The Short Run

In order to make omega theory falsifiable, an omega model must be constructed, which can be seen as constructing a particular stage in which social actors will play their roles. The following auxiliary assumptions are introduced for that purpose:

- There are two social groups: capitalists who are endowed with stocks of physical capital and human capital, and workers with stocks of human

capital and cash balances. Wages are paid at the end of the production period; consequently, workers need to hold cash balances for transaction purposes. Government behavior consists of supplying money to the economic process.

- There are two production sectors: the capitalist sector producing with hired workers and the subsistence sector, in which workers are self-employed. The discipline device in the labor market is based on the gap between the wage rate and the marginal income in the self-employment sector, also called subsistence sector.
- Four markets will constitute the market system: labor, commodity, money, and foreign exchange. Human capital is the same for all, so there is only one labor market. One sole good is produced in society, called good B. Workers and capitalist firms seek to exchange labor services and goods in the labor and commodity markets. Capitalist firms seek to exchange goods in international markets, using foreign exchange as the means of payment. There is no market for renting capital services.
- Capitalists, workers, and the government interact and compete in the market to obtain their selfish objectives, but none has the power to set prices (a perfect competitive market); they all face costless information on market prices and technology.
- The economic process is static and open.

The economic structure of the omega economy consists of a capitalist sector and a subsistence sector; hence, there are two types of production units, capitalist firms and subsistence units in which workers are self-employed. Production technology in the capitalist sector uses capital and labor, whereas in the subsistence sector it uses only labor. The labor productivity level of the subsistence sector is, therefore, lower than that of the capitalist sector; moreover, this productivity level is too low to allow the unit to operate as a capitalist firm (to pay wages and generate profits), and it is called subsistence sector in this sense. In a world of one good, output in the subsistence sector will be destined to direct consumption by the producer, not to market exchange.

If the subsistence sector produced several goods, this assumption would be equivalent to saying that the demand for the goods produced in the subsistence sector comes from the same sector. There would be market exchange, but within the subsistence sector only. Market exchange of goods between the capitalist sector and the subsistence sector will be ignored.

How do capitalist firms extract work effort from workers in an overpopulated society? What is the device to discipline workers? The rate of unemployment cannot be used as the disciplinary device because excess supply of labor is too large. Firms must use a different device.

The Lewis Model: Horizontal Labor Supply Curve

Economist Arthur Lewis (1954) developed a model for an overpopulated society, which is going to be presented here as the first model of the omega theory. He introduced the assumption that firms will pay a premium above the income in the subsistence sector as a device to get labor discipline; that is, in order to extract effort from workers, firms will seek to pay a wage rate that is above the wage earners' opportunity cost, which is equal to the income that they can make in the subsistence sector. This gap will create a cost to wage earners if they are dismissed from the firm due to shirking behavior. The value of the premium is exogenously given.

The Lewis model also assumed that the average productivity of labor in the subsistence sector is constant, which implies that average productivity and marginal productivity of labor are equal. The real wage rate is equal to a premium (say 30%) above the constant average productivity of labor in the subsistence sector; therefore, the real wage rate in the labor market is exogenously determined. If the subsistence sector is constituted by peasants, the model will say that the average productivity of the peasant economy determines the real wage rate of workers in the urban labor market. Hence, the real wage rate will rise if and only if the labor productivity in the peasant economy increases.

Given the real wage rate, capitalist firms will seek to hire workers in a quantity that makes profits the largest. The quantity of employment in the labor market is determined by demand. Due to overpopulation, at this real wage rate, there will be excess supply of labor, which will be self-employed in the subsistence sector. This is the labor market equilibrium in this model.

The Lewis model generates empirical predictions that are refuted by the empirical regularities listed in Chapter 2. First, this model predicts zero unemployment in the short run. This prediction is inconsistent with the reality of the Third World countries, Fact 2, where we observe unemployment, in addition to a significant amount of self-employment. Second, in the long run, the model predicts that real wages cannot rise if capital stock increases. This prediction is inconsistent with Fact 5. Third, real wage and employment are both determined in the labor market by real variables alone, such as capital stock, technology and labor productivity in the subsistence sector. Nominal variables, such as money supply, do not play any role. If incorporated into a general equilibrium system, this labor market model would predict money neutrality, which is inconsistent with Fact 4: nominal variables are correlated with real variables.

A New Model: Rising Labor Supply Curve

The new model will abandon the auxiliary assumption of constant average productivity of labor in the subsistence sector and will assume diminishing returns instead. Production units in the subsistence sector have different labor productivity levels due to differentiated entrepreneurial talents and access to public goods. In this case diminishing returns arise not because of the problem of overcrowding a fixed production factor, such as overcrowding with labor the existing capital stock (there is no capital here), but because of quality differences between production units. Additional units entering into the subsistence sector

will have to work in locations of lower quality, such as distance to public goods (infrastructure). The model assumes the *Ricardian diminishing returns* in the subsistence sector, which is the result of using locations of lower quality as more units enter into the subsistence sector. Figure 4.1(b) can represent this property of Ricardian diminishing returns in the subsistence sector if the curve refers to subsistence units, instead of capitalist firms; and the area under the marginal productivity curve (total income) is equal to labor income, instead of profits and wages.

The implication of this assumption is that marginal productivity of labor in the subsistence sector will be diminishing. Additional self-employed workers will add smaller and smaller quantities to total output in the subsistence sector; consequently, average productivity of labor will also decline as more self-employed workers enter into the subsistence sector.

Workers seek to maximize total income. At the given market wage rate, some workers would be able to make higher income in the subsistence sector, but others would not; hence, the first group will seek self-employment, whereas the second will seek wage employment. The equilibrium allocation of total labor into the two sectors would be when the marginal productivity of labor in the subsistence equals the real wage rate. If the real wage rate were higher, then the second group of workers unable to make higher income in the self-employment sector would increase, which implies an increase in the number of workers seeking wage employment. Thus we have just derived the labor supply curve: the higher the wage rate, the higher the quantity of labor supplied to the labor market. The marginal productivity curve of the subsistence sector represents, at the same time, the labor supply curve.

Capitalist firms will still seek to pay a wage rate that is higher than the opportunity cost of labor. This gap is a premium that intends to be the discipline device. The wage premium will be applied to the rising labor supply curve. When shifted upwards by the proportion of the premium, the labor supply curve will then become the effort extraction curve. The effort extraction curve will operate as a restriction for wage determination in the labor market.

The selfish motivations of capitalists and workers that were assumed in the epsilon society will also be applied to the omega society. However, those motivations will be pursued under different constraints.

The behavior of capitalists

Capitalists seek profit maximization. For the typical firm, this motivation and the constraints can be written as in equation (4.7) in Chapter 4, which is re-written here just for convenience:

$$\begin{array}{ll}
 \text{Maximize} & P' = P_b Q_b - P_c D_c - P_h D_h \\
 \text{Subject to} & \\
 & Q_b = \varphi(D_h, K_b), \text{ where } \lambda^* = 1 \\
 & Q_b = D_c / c \\
 & X_b = (1/z^*) D_c
 \end{array} \tag{5.1}$$

This is the same set of structural equations that were established in the epsilon model (Chapter 4). The equilibrium condition of the individual firm is therefore similar to the condition that was obtained in the epsilon model; namely, the market nominal wage must be equal to the value of the net marginal productivity of labor in the firm, in which “net” means total output minus the output utilized to pay for the required imported inputs, the good C. This equilibrium is (trivially) stable.

The comparative statics at the individual firm level and then the further aggregation over all firms will be similar to the procedure shown in equations (4.8) and (4.10). The latter is reproduced now as system (5.2):

$$D_h = H(P_b, P_c, P_h, K_b), H_1 > 0, H_2 < 0, H_3 < 0, H_4 > 0 \quad (5.2)$$

Subject to

$$P_b S_b = (P_c D_c - P_b X_b) + P_h D_h$$

$$X_b = (1/z^*) D_c$$

$$P_h \geq P_h^*,$$

$$w \geq w^* = (1 + p) v' \quad (5.3)$$

The first equation is the aggregate labor demand function. The constraints include the aggregate budget constraint of firms, the aggregate export function of firms, and nominal wage downward stickiness. Equation (5.3) indicates the new constraint: the condition that the market real wage rate must be equal to or higher than the threshold of efficiency wages (w^*), which must in turn be equal to the wage rate that is determined at the point where the effort extraction curve (different from the labor supply curve) crosses the labor demand curve.

The behavior of workers

Workers are endowed with human capital and cash balances. They seek to maximize total real income subject to their resource endowments and the rationing mechanisms of the labor market.

As a first option, workers seek employment in the capitalist sector, for the market wage rate is higher than the income they can make in the subsistence sector. The incentive system leads them to this endeavor, but jobs for all workers are not available, no matter how hard each worker seeks wage employment. Because there is a turnover of workers in the capitalist sector due to dismissals of those workers found shirking, there is a probability of finding a job in the capitalist sector. Assume this probability is the same for all workers.

Let w^e represent the workers' expected wage if seeking a job. For an exogenously given probability π of finding a job, the expected wage depends positively on the market wage rate. These assumptions imply a uniform expected wage for all workers, which can be written as

$$w^e = \pi w, \text{ where } 0 < \pi < 1 \quad (5.4)$$

As the second best solution, the workers who become excluded from wage employment will choose between unemployment and self-employment. These workers will evaluate the expected wage (what they would get after seeking and finding a job) against the sure income they can make in the subsistence economy.

For a given market real wage rate, there will be an expected income if unemployed. For some workers the expected wage rate would be higher than the income they can make in the subsistence sector and for others it would be lower. The first group would seek jobs in the labor market, whereas the second group would choose self-employment. The equilibrium allocation of workers between unemployment and self-employment will take place when the expected wage rate is equal to the marginal productivity of labor in the subsistence sector.

Because workers first seek wage employment, the behavior of *wage earners* will be similar to the behavior presented in the epsilon model, equations (4.4), (4.5), and (4.9). The latter is reproduced here just for convenience

$$\begin{aligned} D_b &= F(P_b, D_h, P_h), \quad F_1 < 0, F_2 > 0, F_3 > 0 \\ D_m &= P_b L(D_b), L' > 0, L'' < 0 \\ \text{Subject to the aggregate budget constraint of wage earners} \\ P_h D_h &= P_b D_b + (D_m - S_m) \end{aligned} \quad (5.5)$$

5.3 General Equilibrium

The capitalist sector

There are four markets that operate under perfect competition in the capitalist sector. The conditions of equilibrium for each market are

$$\begin{aligned} \text{Labor market} & \quad S_h \equiv D_h + E_h, P_h \geq P_h^*, w \geq w^* = (1+p)v' \\ \text{Commodity market} & \quad S_b = D_b \\ \text{Foreign exchange market} & \quad P_c^* D_c = P_b^* X_b \\ \text{Money market} & \quad S_m = D_m \\ \text{Subject to the aggregate budget constraint} \\ & \quad P_b (D_b - S_b) + P_c (P_c^* D_c - P_b^* X_b) + (D_m - S_m) = 0 \end{aligned} \quad (5.6)$$

The labor market is non-Walrasian; it operates with excess labor supply E_h , and subject to the constraint that the nominal wage cannot fall and the real wage must lie within the range of efficiency wages. The rest are Walrasian markets.

The market system will solve for the values of the endogenous variables, for given values of the exogenous variables. These variables are

Endogenous variables: $P_e, D_h, P_b, P_h, w, E_h, Q_b, Y, W, P$

Exogenous variables: $z^*, r^*, P_h^*, K_b, S_h, \delta, S_m$

The set of endogenous variables includes some variables that were exogenous in the micro level of analysis or partial level of analysis, such as nominal prices. This is in accord with the principle of increasing endogenization of the aggregation process.

It should be noted that the general equilibrium conditions in the capitalist sector are equal to the ones established for the epsilon model (Chapter 4). The core of the general equilibrium model will also be the same—equation (4.12). As before, equilibrium in the foreign exchange market is incorporated into the labor market equation. That leaves us with two Walrasian markets (good B and money), of which one can be eliminated due to Walras' Law. Then general equilibrium can be established by solving two markets, say the money market and the labor market. The equations of the core (4.12) are replicated here for easy reference. They are

$$D_h = H(P_e; P_h, z^*, K_b), H_1 > 0, H_2 < 0, H_3 > 0, H_4 > 0 \quad (5.7)$$

$$S_m = P_e P_b^* L(W) = M(P_e, D_h; P_h, P_b^*), \text{ where } M_i > 0, \text{ for all } i$$

Subject to

$$D_h < D_h^* < S_h$$

The constraints indicate that we are searching for the existence of general equilibrium with excess labor supply. In this system, there are two equations and two endogenous variables, the exchange rate P_e and the employment level D_h . The solution is simultaneous. Once these values are known, the rest of the endogenous variables can be solved just by implications.

Figure 5.1 depicts the core of the general equilibrium. Panel (a) shows the labor market, in which the demand curve h is presented in terms of the nominal wage rate. The position of this curve is determined once the exchange rate is known. Panel (b) shows the money market, in which the particular position of the demand curve M is determined once the wage employment level is known. Hence, there is a pair of values that determine the equilibrium values of these endogenous variables. The solution is simultaneous, as shown by point e and point F in Figure 5.1, panels (a) and (b). Once the exchange rate is known, the price level is determined, so is the real wage rate of equilibrium.

The equilibrium of the labor market in terms of real wage rates is represented in Figure 5.1(c). The curve nm represents the labor supply curve and, given the wage premium p , we derive the effort extraction curve $n^* m^*$. The curve HR represents the labor demand curve. The market real wage rate must belong to the set of efficiency wages; thus there exists a real wage w^* that is the minimum value that the real wage can take in order to maintain the labor demand curve fixed (the level of the marginal productivity of labor). Graphically, this minimum wage is determined by the point at which the effort extraction curve and the labor demand curve cross each other: point g . The value of w^* is the threshold of efficiency wages and determines the maximum wage employment level D_h^* . Point E indicates the equilibrium position (the counterpart of point e in panel (a)). The equilibrium employment level is OA and the equilibrium real wage is w^0 . The excess labor supply is equal to $A0'$, where $00'$ is the total labor supply.

The subsistence sector

Once the values of equilibrium in the labor market have been determined, the allocation of the excess labor supply into unemployment and self-employment will be solved. The relevant equations to solve this allocation include

$$\begin{aligned} E_h^0 &= L_s + U \\ V &= J(L_s), \text{ where } J' > 0, J'' < 0 \\ \pi w^0 &= J'(L_s) \equiv v', \quad 0 < \pi < 1 \end{aligned} \tag{5.8}$$

The first equation points out that in the omega economy excess labor supply includes unemployment and self-employment. (This is different from the epsilon economy, where excess labor supply takes the form of unemployment only.) The second equation is the production function in the subsistence sector, which is subject to Ricardian diminishing returns. These two are the structural equations. The third equation shows the equilibrium condition for the allocation of labor to self-employment: the marginal productivity of labor in the subsistence sector (v') must be equal to the expected wage (πw^0).

Given the curve of the marginal productivity of labor in the subsistence sector, the self-employment level will be determined by the expected wage rate; given the total amount of excess labor supply, the unemployment level of equilibrium will be determined once the self-employment level is determined. Thus, the unemployment level is a residual of the residual labor; it could even be zero in equilibrium.

The static general equilibrium solution of the omega model is clearly sequential. First, equilibrium in the capitalist sector is determined, which determines the excess labor supply as well; then, equilibrium self-employment in the subsistence sector is determined, which determines the amount of unemployment.

Figure 5.1 also shows the general equilibrium of the omega economy. Once the capitalist sector solution is obtained, by the interactions of the labor market and the money market, shown in panels (a) and (b), wage employment level and the real wage rate are known, as shown in panel (c). Then the expected wage is determined. This is marked as a fraction of the market wage rate. This value must be equal to the marginal productivity of labor, which occurs at point f; hence the self-employment level is determined. Unemployment is the residual of the total excess labor supply. Thus, the allocation of the excess labor ($A0'$) to unemployment (AB) and self-employment ($B0'$) is solved. In order to assure labor discipline and keep the curve HR unchanged, the gap between the real wage and the marginal productivity of the self-employed workers must be equal to or higher than the wage premium. At point B , the equilibrium real wage is indeed above the curve n^*m^* . General equilibrium is thus obtained.

As can be seen in Figure 5.1(c), the higher the value of π or p , the less likely is to satisfy the efficiency wage condition $w \geq (1+p)v'$; hence, in order to have a stable system, these parameters cannot take independent values. The values of these parameters that are consistent with the efficiency wage assumption can be obtained from the following conditions:

$$\begin{aligned}
w &\geq (1+p) v' \\
\pi w &= v' \\
\pi (1 + p) &\leq 1
\end{aligned} \tag{5.9}$$

The first equation is the efficiency wage condition; the second is the equilibrium condition in the subsistence sector; and the third is derived from the others, which indicates that stability of the general equilibrium requires a particular relationship (an equilateral hyperbola) between the probability of finding wage employment and the wage premium paid to wage earners. Given this stability condition, the general equilibrium can indeed be solved sequentially, as shown above.

For example, the system will be stable if $\pi=0.80$ and $(1+p)=1.25$, which means that 80% of real wage can be earned if seeking jobs and finding a job (equivalent to two months of unemployment in a year) and the market wage rate is 25% above the opportunity cost of labor; if $\pi=0.70$ (equivalent to roughly three months of unemployment) and $(1+p)=1.30$, the system will also be stable. But if $\pi=0.50$ (six months of unemployment in a year) and $(1+p)=1.30$, the system will be unstable. The model assumes values that generate stability.

The labor market is clearly a non-Walrasian market. Workers are willing to exchange their labor at the market real wage, but not all are able to do so. Unemployment or self-employment is the result of economic choice made by those workers excluded from the labor market. In the latter case, unemployment is “voluntary” in the sense that it is the result of economic choice, for they could take the low self-employment income. But it is involuntary in the sense that unemployment is not the most preferred situation for workers. Unemployment or self-employment is an alternative under a second-best situation. Because self-employment generates income that is below the market wage rate for the same skill endowments, self-employment may be called *underemployment*.

As can be seen in Figure 5.1(c), the equilibrium rate of unemployment need not be a positive number; it could be zero. In omega, a labor abundant society, unemployment is not a necessity for the functioning of the capitalist system; it is underemployment the variable that plays that role. Underemployment is necessary for labor discipline and thus for high labor productivity and high profits in the capitalist sector. For workers that are not needed in the capitalist sector, self-employment is a way to generate their own income and make viable the capitalist sector in an overpopulated society. Underemployment, not unemployment, is a necessity for the functioning of capitalism in an overpopulated society.

Production and distribution

National income (Y) of equilibrium and its distribution between social groups can be represented as follows:

$$\begin{aligned}
Y &= P + W + V = P + w D_h + v L_s \\
S_h &= D_h + L_s + U
\end{aligned} \tag{5.10}$$

The term v measures the average productivity of labor in the subsistence sector.

In Figure 5.1, panel (c), the national income of equilibrium is equal to the aggregation of two areas: the area under the curve HE (total output in the capitalist sector) plus the area under the curve mf (total output in the subsistence sector). Capitalist firms and the small units in the subsistence sector produce the same good. In this world of one commodity, this aggregation is simple and fully justified.

The static general equilibrium solution is clearly inefficient. The allocation of labor given by the intersection of the supply and demand curves would maximize national income. This equilibrium occurs at point N in Figure 5.1(c), where the labor supply curve is given by the curve nm, measured from origin O. But that solution is not economically attainable. It would require a Walrasian labor market. The omega economy is inherently inefficient.

The equilibrium income distribution can also be seen in Figure 5.1(c). The area under the curve HE is the total output generated in the capitalist sector, which is distributed between profits and wages. The area under the curve mf is the total output generated in the subsistence sector. The functional income distribution is thus determined. To determine the personal income distribution, we must take into account the hierarchy of income s among workers: the wage rate is higher than the average income of the self-employed. The unemployed has an expected wage that is equal to the marginal income in the subsistence sector.

Among the Walrasian markets of the system, the money market is explicitly in equilibrium. The foreign exchange market is implicitly in equilibrium because it is incorporated into the labor demand curve. It can be shown that the commodity market is also in equilibrium. The quantity of good B supplied to the domestic market is given by the difference between total net output (net of exports to pay for the technological required inputs) and profits; in Figure 5.1(c), this is equal to the area of the wage bill, which in turn is equal to the quantity demanded. Thus, in fact, there exists general equilibrium in the market system.

Could the inequality among workers endowed with the same human capital persist? Those workers who are excluded from the labor market make up the poor groups of society. Could they offer their labor at lower nominal wages to get wage employment? No, they could not, because social norms impede this type of competition among workers. Furthermore, there is a limit to the level of wage employment given by real wage w^* . Therefore, workers who are excluded from the labor market have no choice but to seek wage employment (the unemployed) or remained self-employed (the underemployed).

Workers who are excluded from the labor market could also seek to set up firms or expand the size of their current productive units where they are self-employed. Skill level does not present a barrier. They could rent capital to open a firm, but there is no market for capital service. They would have to buy physical capital, which in turn needs financing. But these workers are also excluded from the credit market and the insurance market, making these projects unviable. (The theories of exclusion of credit and insurance markets are

presented in Chapter 8 below.) The general equilibrium shown above is indeed an equilibrium situation: no one has both the power and the will to change it.

In sum, the static general equilibrium of the omega economy is determined sequentially: first, the equilibrium in the capitalist subsystem is attained; then output and employment in the subsistence sector is determined; unemployment is a residual of the excess labor supply. Unemployment and underemployment form the excess labor supply. In the omega society, the incentive system to get the optimal work intensity, which implies the highest feasible profits, is based on the existence and persistence of underemployment; by contrast, in the epsilon economy it is based on unemployment. The workers who are excluded from the labor market do not receive income from any insurance program via taxes; they make the system viable by generating incomes by themselves. This is the role played by the subsistence sector; without the subsistence sector, the capitalist sector would not be socially viable in an overpopulated society. In the static system, the values of general equilibrium will be repeated period after period, as long as the exogenous variables of the system remain unchanged.

5.4 Beta Propositions

The sequential solution of the general equilibrium in omega society implies that the empirical predictions of the model can be analyzed for the two sectors independently. If the market equilibrium that has been attained in the capitalist sector is stable, comparative statics can be applied to derive beta propositions. It was shown above that the core of the market equilibrium is indeed stable. Comparative statics can thus be applied to this equilibrium in order to derive beta propositions: the effects of exogenous variables upon endogenous variables. The endogenous and exogenous variables in the capitalist sector of the omega model are the same of those of the epsilon model, so are the structural equations; therefore, it is expected that the effects of exogenous variables upon endogenous variables in the capitalist sector of the omega model would be similar to those presented for the case of the epsilon model.

In the subsistence sector, there exists also stability, as shown above; then comparative statics can be applied here. The variables are

Endogenous variables: L_s , v , V

Exogenous variables: w , E_h

Real wage and the excess labor supply are endogenous variables in the omega capitalist sector but are exogenous variables in the subsistence sector. This is just the indication that the general equilibrium is determined sequentially. The endogenous variables include the self-employment level and average and total income in the subsistence sector. Unemployment is residual. The comparative statics of the general equilibrium model is now applied to obtain the empirical predictions of the model for the short run.

Firstly, consider the effect of government behavior. An increase in the money supply generates directly an excess of cash balances willingly held by workers, who will then intend to restore the equilibrium by using this excess to buy more quantities of good B in the period of adjustment. This adjustment would reduce both the exports and foreign exchange inflows and thus an increase in the exchange rate. Then the price level would increase and the real wage would fall; hence, firms would seek to hire more workers (from the pull of excess supply) and produce more output to obtain more profits. This would be the first round effect. The second round effect will be ambiguous because the fall in the real wage rate and the increase of employment imply an ambiguous change in the real wage bill, which in turn imply an ambiguous and small change in the demand for real cash balances and on the price level. Therefore, the second round effect upon the price level would be small and may be neglected. The effect of an increase in money supply is to increase employment by decreasing the real wage rate.

The effect of an expansionary monetary policy can be seen in Figure 5.1(c). Point E represents the initial general equilibrium. An increase in the money supply increases the price level, which implies a fall in the real wage rate and an increase in the employment level. The initial equilibrium situation, at point E, will be moved to another equilibrium situation, which will lie below point E, in the segment E-g. The government can then reduce the excess supply of labor using monetary policy. A sufficient large increase in money supply will ultimately reach the effective full employment level, at point g. If money supply keeps increasing, employment will remain fixed, and only the exchange rate and consequently the price level will rise, which will then induce increases in the nominal wage rate in the same proportion in order to avoid the fall below the minimum efficiency wage rate (w^*); that is, *money becomes neutral*. It should be clear that the Walrasian market equilibrium, point N, is unattainable in the omega model.

In the short run analysis, the relevant exogenous variables are the international terms of trade and the international interest rate, as in the case of the epsilon model. The basic endogenous variables are total output and its distribution between social groups. In the omega model there are two sectors: capitalist and subsistence, and there are three social groups in the distribution process: capitalists, wage-earner workers and self-employed workers. Unemployed workers receive no income in the period of analysis.

Consider the effects of an increase in the international terms of trade. As can be visualized in Figure 5.1(c), an increase in the international terms of trade (z^*) shifts the labor demand curve HR outwards. The initial general equilibrium is at point E. At the current real wage rate, firms would seek to hire more workers (drawn from the pool of excess supply); hence, the wage bill would increase. Given the quantity of money supplied, the build-up of cash balances would imply a fall in the quantity of good B demanded, which would increase both exports and the inflow of foreign exchange, which would lead to a fall in the exchange rate, which in turn would lead to a fall in the domestic price level and a rise in the real wage rate; hence employment would tend to fall, but not in a magnitude that would eliminate the initial increase. As a final result, both employment and real wage rate will increase. The initial equilibrium situation, given by point E, will be moved to another equilibrium situation, say point E (not marked), which will be located to the north-east of point E.

The omega model assumes that the supply of money is exogenous, but in a particular sense: the government reacts to changes in output due to external shocks, seeking to alter these results in the direction indicated by the logic of maximization of votes. Money supply is the only instrument that the government can manage to achieve this objective. As in the epsilon model, money supply is defined as quasi-exogenous, and its effect will alter only partially the initial effect of the international terms of trade. Therefore, the net effect (net of government reaction) of a change in the international terms of trade found above will prevail.

What is the subsequent effect of changes in the international terms of trade upon the subsistence sector? The effect is indirect, through the effect upon the labor market. The increase in the international terms of trade implies an increase in both the real wage rate and wage employment, which in turn implies a reduction in the excess labor supply, and a fall in unemployment. However, the increase in the real wage rate leads to higher expected wage rate and thus to a decrease in self-employment and to a consequent increase in unemployment. The final outcome is the decrease in self-employment together with an ambiguous change in unemployment.

The effect of an increase in the international terms of trade upon income inequality can also be visualized in Figure 5.1(c). The new equilibrium in the capitalist sector indicates higher output level and higher wage bill, which implies an ambiguous change in profits in the capitalist sector. Total income in the subsistence sector falls. Regarding functional distribution, the overall change is ambiguous, in spite of the fact that the total income in the subsistence sector falls. Concerning, personal income distribution, both the real wage rate and the mean real income in the subsistence sector increase, but the mean income of capitalists is ambiguous; hence the change in the personal distribution is also ambiguous.

The effect of the international interest rate upon production and distribution in the short run in the omega society is just the opposite of the international terms of trade. An increase in the international interest rate, maintaining the domestic interest rate fixed, causes an outflow of foreign exchange, which results in the rise of the exchange rate and then in the price level. The result is a fall in the real wage rate and a fall in employment in the capitalist sector. The effects upon the other endogenous variables are determined just by implications.

The final exogenous variable is the initial inequality. Changes in the initial inequality would occur if the individual endowment of economic assets is redistributed exogenously. An increase in this variable implies a higher degree of concentration of economic assets. This would occur if the capitalist class becomes smaller in size or, given its size, the concentration increases within the class. The effect of an increase in the initial inequality on output will be nil because the production efficiency of the economy is independent of the degree of income inequality (an assumption to be relaxed later on). The allocation of labor to the capitalist sector and to the subsistence sector depends on real wages, which will not change as a result of a higher concentration of assets. With regard to inequality, the functional distribution will remain unchanged, but personal distribution will become more unequal, for the same amount of profits will become concentrated in fewer individuals. The effect of an increase in the initial inequality on the degree of income

inequality will be positive: The higher the inequality in the distribution of the *stock* of resources, the higher the inequality in the distribution of the *flow* of incomes.

In sum, from the omega model we have been able to derive the reduced form equations for the level of national income (Y) and the degree of inequality in its distribution (D), which are:

$$Y^0 = F^\omega(z^*, r^*, \delta; K_b, S_h) \quad (5.11)$$

$$F_1 > 0, F_2 < 0, F_3 = 0, \text{ and } F_4 > 0, F_5 = 0$$

$$D^0 = G^\omega(z^*, r^*, \delta; K_b, S_h) \quad (5.12)$$

$$G_3 > 0 \text{ and } G_i = ? \text{ Otherwise}$$

Equation (5.11) indicates that factor endowments determine the level of national income, with short run variations due to the effect of changes in two exogenous variables: the international terms of trade and the international interest rate. Similarly, equation (5.12) indicates that the initial inequality determines the level of inequality in the distribution of national income, with short run variations around that level due to changes in the other exogenous variables. The beta propositions are given by the signs of the partial derivatives. (The mathematical proof is shown in Appendix A.)

The omega model shows, by construction, that the macro behavior of output and distribution has micro foundations and that, at the same time, the underlying micro behavior of social actors has macro foundations. The unity of knowledge is thus assured.

5.5 Empirical Consistency: Third World Countries with Weak Colonial Legacy

Do Third World countries resemble the omega economy? Are the beta propositions of this omega model consistent with the empirical regularities listed in Chapter 2?

The omega static model can be confronted against the relevant regularities, which include Facts 2 and 4. The static model indeed predicts Fact 4, the interplay between nominal and real variables in the short run.

Regarding Fact 2, the omega model predicts that any general equilibrium situation will imply that the average wage rate is higher than the average income of the self-employed for similar level of skills ($w > v$). This equilibrium condition is observable and it thus makes the model refutable. The model also predicts the existence of unemployment. In sum, the empirical predictions of the omega model consistent with Facts 2 and 4.

According to the omega model, labor discipline in the capitalist sector is ensured via underemployment, not via unemployment. Sigma society may operate with or without unemployment; it is not a necessity, as it was in the epsilon society. Therefore, underemployment is a necessity for the functioning of capitalism in an overpopulated society. The peasantry thus plays a significant role in the functioning of capitalism in the Third World. If for some reason these lands became unproductive, the landless peasantry

would increase the size of unemployed; then capitalism could not operate in this overpopulated society with a very large unemployment rate.

Consistent with the regularity established under Fact 2, the proportion of workers engaged in wage employment should be higher in the First World compared to the Third World. A sample of eight First World countries and 10 Latin American countries in the ILO dataset showed, indeed, that the ratio of wage employment to total labor force had a mean value of 84% in the First World and 59% in Latin America around 1996.⁶

According to the omega model, the size of the excess labor supply is equal to the sum of unemployment plus underemployment. The latter is defined as the self-employed workers earning incomes below the market wage rate, for similar skills. If this gap did not hold in reality, the omega model would fail.

Empirical studies measuring this income gap are not frequent in the international literature. Only two country studies can be cited at this moment. A study on Peru, based on a national household survey of 2003, found, firstly, that indeed the wage rate was higher than the mean income of the self-employed: it was 80% higher among workers with primary level education (low skill), and 30% higher among workers with secondary level education (high skill); and secondly, that the rate of underemployment was 51% and that of unemployment was 7%, estimating a rate of total excess labor supply at 58% (Figuerola, 2010, Tables 4 and 5, pp. 124-125).

The other study is on Brazil. The study is based on a sample of the national census of 1980, in which the sample size was 3% and the universe the large cities of Brazil and workers with 11 years of education or less (excluding post-secondary levels). This study finds that the income gap between wage earners and the self-employed (the “unprotected”) is around 30% (Telles, 1993, Table 1, p. 239). No estimation of the rate of excess labor supply is made. Thus, according to the available data, Fact 2 seems to be consistent with the predictions of the omega model.

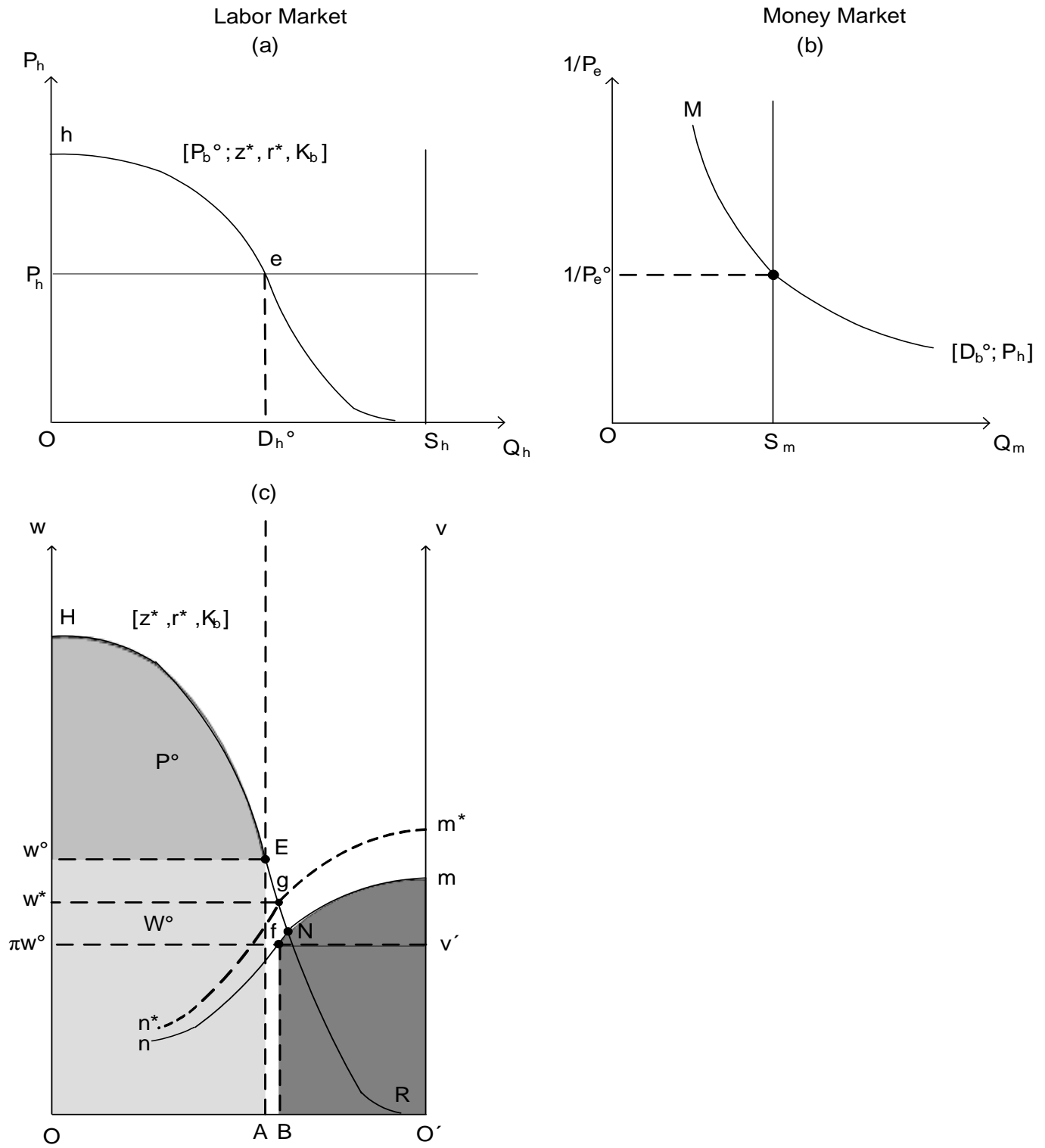
The common practice of using the unemployment rate as the criterion for making international comparisons in the excess labor supply is thus not warranted. The rate of excess labor supply takes two forms, unemployment and underemployment, in the Third World countries. It is clear that the excess labor supply in the First World is given by unemployment alone. But in the Third World unemployment figures largely underestimate the magnitude of the total excess labor supply.

Two most notable facts of the Third World countries seem to be consistent with the predictions of the omega model. Indeed, Facts 2 and 4, stated in Chapter 2, are predicted by the omega model. However, there is another empirical regularity that is also notable in the Third World. This is Fact 3: the existence and persistence of income gaps between ethnic groups. Omega theory makes abstraction of ethnic differences in society and could not

⁶ The sample includes Japan, Denmark, Germany, the United States, France, Sweden, Spain and Portugal for the First World, and Mexico, Brazil, Panama, Costa Rica, Peru, Colombia, El Salvador, the Dominican Republic, Ecuador, and Bolivia for Latin America. (ILO, *World Employment Report 1997*, Tables 2A, 2D, and 2E). (Calculations made by the author).

explain this phenomenon. Another theory of the Third World is then needed. This is presented in the next chapter.

Figure 5.1. General Equilibrium in the Omega Society



CHAPTER 6

THE THIRD WORLD WITH COLONIAL LEGACY

In epsilon and omega societies, individuals are homogeneous in every respect, except in their endowments of economic assets. Social assets are equally distributed between individuals; hence, these societies are socially homogeneous. An abstract heterogeneous and hierarchical society, which will be called sigma society, will be presented in this chapter. It aims at explaining production and distribution in the Third World countries with colonial legacy.

6.1 Sigma: Socially Hierarchical and Overpopulated Society

Sigma is also a class society. As in the case of omega, it is also an overpopulated society. But contrary to omega, individuals are not entitled with unequal political rights and duties. Sigma is not only a class society, but it is a socially heterogeneous society. Sigma is thus a hierarchical society.

Conceptually, assets are goods that provide a flow of incomes. They include not only economic assets (land, physical capital, and human capital), but also social assets; therefore, the assumption here is that social assets also generate a flow of incomes. Social factors are thus introduced into the economic process. However, social assets are special goods for they belong to the realm of rights and entitlements granted to individuals in society. They are no physical goods, nor are they marketable.

Social assets basically refer to political and cultural assets. Political assets are defined as the capacity of individuals to exercise individual and collective rights, including the right to have rights. Inequality in the individual endowment of political assets generates a hierarchy of citizens in society, first-class and second-class citizens. As a result, not all individuals are equal before the law; moreover, not all individuals have the same degree of access to public goods supplied by the state.

Cultural assets can be defined as the right of social groups to cultural diversity in a multicultural and multiethnic society. Inequality in the endowment of cultural rights generates ethnic groups with a hierarchy of ethnic markers in society: there are first-class and second-class races, languages, religions, and customs. These markers are called cultural because their hierarchy are socially constructed and are also transmitted from generation to generation. Inequality in cultural assets leads to social practices of segregation, exclusion, and discrimination against some ethnic groups.

Individual endowments in political and cultural assets are assumed to be highly correlated in sigma society; hence, only unequal political assets—differences in

citizenship—will be included in the construction of sigma society. Sigma society is not only socially heterogeneous; it is hierarchical. There exist social classes and under-classes.

The primary assumptions of the sigma theory (σ) are summarized as a set of alpha propositions as follows:

$\alpha(\sigma).(1)$ *Institutional Context*: (a) Rules: Individuals participating in the economic process are endowed with economic and political assets. Economic assets are subject to private property rights. Individuals exchange goods subject to the norms of market exchange, which include the norm that nominal wages in labor markets cannot fall. The political system includes formal or informal norms for excluding some social groups from full citizenship; (b) Organizations include households, firms, and the government.

$\alpha(\sigma).(2)$ *Initial Conditions*: (a) Initial inequality: individuals are endowed with unequal quantities of economic assets and unequal political entitlements. The social structure includes two types of social classes: capitalists and workers and also first and second class citizens (the under-class). (b) Initial factor endowments: the stock of capital per worker is such that the marginal productivity of labor is zero, that is, sigma is an overpopulated society.

$\alpha(\sigma).(3)$ *Economic Rationality*: Individuals act guided by the motivation of self-interest. Capitalists seek two objectives: the maintenance of social position and the maximization of profit, such that the former objective has priority. In the labor market, workers seek to maximize wages and minimize effort; hence, capitalists use devices to extract effort from workers.

6.2 A Static Model of the Sigma Theory: The Short Run

In order to make sigma theory falsifiable, a sigma model must be constructed, which can be seen as a task of constructing a particular stage in which social actors will play their roles. The following auxiliary assumptions are introduced for that purpose:

- The existence of a second-class citizenship in sigma society comes from the legacy of its colonial history. The most important legacy of colonial systems is political inequality, which generates a social hierarchy between the descendants of conquerors and the descendants of dominated populations.
- There are four social actors: capitalists who are endowed with stocks of physical capital, high-level human capital, and first-class citizenship; two groups of workers, one endowed with high-level human capital and first-class citizenship, and the other endowed with low-level human capital and second-class citizenship; and the government.
- The economic structure is composed of three sectors: the capitalist sector, a subsistence sector in which x-workers can generate income as self-employed, and another subsistence sector in which z-workers can generate income as self-employed. Output produced in each subsistence sector is not

for market exchange, just for consumption within the sector. One sole good is produced in society, called good B.

- Four markets will constitute the market system: labor, commodity, money, and foreign exchange. There is only one labor market, for x-workers alone, in which there is overpopulation. Money supplied by the government is used as the means of payment. Capitalist firms seek to exchange goods in international markets, using foreign exchange as the means of payment. Wages are paid at the end of the production period; consequently, workers need to hold cash balances for transaction purposes.
- Capitalists, workers, and the government interact and compete in the market to achieve their selfish objectives, but none has the power to set prices (a perfect competitive market); they all face costless information on market prices and technology.
- The economic process is static and open.

Table 6.1 presents the social structure of sigma in a matrix form. There are three ethnic groups: the Blues, the Reds, and the Purples, such that the purples are the result of miscegenation of the other two races. Racially, the capitalist class is blue, x-workers are purple, and z-workers are red.

In a paper that analyzes the theoretical relationships between consumer preferences and culture, Akerlof and Kranton (2000) construct an abstract world of two ethnic groups, the Greens and the Reds, in which the Greens are the dominant group. To use primary colors, they will be called Blues and Reds here. As in that paper, it is assumed here that people cannot choose their ethnic identity; ethnicity is exogenous.

The social matrix shows a highly correlated society in the endowment of assets. The Blues are highly endowed with economic and political assets; the Reds are very poor in those endowments; and the Purple lie in between. Three social groups are thus identifiable; for easy reference and for reasons that will become apparent later on, they will be called by the letters A, X, and Z. The results of the model would not change much if we assumed that part of z-workers are endowed with skilled labor and part with unskilled, but the model would be more complex.

Sigma society can now be distinguished analytically from a socially homogeneous capitalist society, such as the epsilon society. If epsilon society were represented by Table 6.1, there would still be two social classes (capitalists and workers) but only one citizenship class (C_1). There would still be three ethnic groups, but only one degree of citizenship for all (C_1). In epsilon society, therefore, racial differences would become unimportant, and the social matrix would collapse into two social groups only: A (capitalists) and X (workers), and the social group Z would not exist. Thus inequality in political assets or citizenship is the essential factor that distinguishes a sigma society from epsilon society. This distinction also applies to omega society.

On factor endowments, the sigma model assumes that there exists overpopulation of x-workers. The labor market operates as a non-Walrasian market, in which equilibrium takes place with excess labor supply. The excess labor supply becomes, in part, self-

employed in small production units in the x-subsistence sector and, in part, unemployed (seeking wage employment and with an expected income).

The model will also assume that a certain threshold of human capital is needed to learn the technology used in the capitalist sector. Z-workers are by assumption endowed with human capital that is below this threshold. Z-workers are out of the labor market and totally self-employed in the z-subsistence sector. The level of labor productivity is highly differentiated between the three sectors: the capitalist sector shows the highest level, the z-sector the lowest, and that of the x-subsistence sector lies in between. This order reflects differences in capital endowments (physical and human) between firms and subsistence units, which also imply differences in levels of technology between them. Labor productivity in the subsistence sectors is too low to pay wages and generate profits.

Production in the subsistence sectors is for direct consumption by the producer rather than for the market. Subsistence units are, therefore, production and consumption units at the same time. This assumption is a logical implication of modeling a one-good economy. In a several goods model, this assumption would imply that exchange takes place within each subsistence sector, based mostly on the rules of reciprocity exchange, and also between sectors based on the rules of market exchange. To be sure, this is not to deny the existence of market exchange between the capitalist sector and the non-capitalist sectors; the assumption of the model is that market exchange between sectors is not the essential factor to understand production and distribution in the Third World, and thus may be ignored. The empirical refutation of the model will tell us whether this is a good assumption or not.

6.3 General Equilibrium

In this model, sigma society operates with three sectors: capitalist, x-subsistence sector, and z-subsistence sector. The model assumes that the z-subsistence sector is totally independent of the other two.

It should then be clear that in this model sigma society can be seen as an omega society—including the capitalist sector and the x-subsistence sector—plus an independent z-subsistence sector. The omega part is just the omega model that was previously presented in Chapter 5. Therefore, we already know a lot about sigma society. We know that the interactions between the capitalist sector and the x-subsistence sector are such that equilibrium is sequential: the capitalist sector is determined firstly and then the x-subsistence sector is residual. The core of the general equilibrium in the sigma model is the same as in the omega model—equation (5.7), shown in Chapter 5.

The sigma model assumes that z-workers could not compete with x-workers in the labor market. Z-workers are endowed with low-level human capital for the technology being used in the capitalist sector. Thus, their human capital endowments are not suitable for wage employment. They are not employable by firms; therefore, they are not part of the labor supply in the labor market. Capitalist firms cannot make profits employing them, as there would be much need to invest in their training, when at the same time x-workers are in excess supply. It is the lack of profitability that lies behind the total exclusion of z-workers from the labor market. Therefore, in this particular sigma model, z-workers are not

part of the working class. They constitute the *underclass*: workers “who are largely expendable from the point of view of the logic of capitalism” (Wright, 1997, p.28).

The only pending task to attain general equilibrium is to specify the characteristics of the z-subsistence sector and its role in the whole society. Assume that production in the z-subsistence sector is subject to the diminishing returns *a la Ricardo*. In comparison with the x-subsistence sector, the average productivity of labor in the z-subsistence sector also declines as more workers participate in the sector; however, the difference is that the level of productivity (the position of the average productivity curve) is lower in the z-subsistence sector. The difference in productivity levels is due to differences in human capital endowments.

The behavior of the z-subsistence sector can be stated as follows:

$$\begin{aligned} v_z' &= \pi_z w_z, 0 < \pi_z < 1 \\ S_{hz} &= D_{hz} + U_{hz} + L_{hz} \end{aligned} \quad (6.1)$$

These are similar conditions to those presented in the x-subsistence sector, if there were a labor market for z-workers. If this were the case, z-workers would allocate their labor between wage employment and self-employment by equalizing the marginal productivity of labor in the z-subsistence sector and the expected market wage rate. However, just for the sake of simplicity, the model will assume that such labor market does not exist ($w_z=0$).

The model also assumes that all z-workers are self-employed independent of the marginal productivity of labor. If it is positive, the sector is composed of individual producers, each getting its marginal productivity as income; if it is zero, the sector has the institutional norm that production and distribution is communal, in which every individual receives the average income of the sector, which is always positive. Hence, the equilibrium condition is $S_{hz}=L_{hz}$. Total output is then determined.

In sum, all z-workers are self-employed in the z-subsistence sector because their human capital endowments are too low to operate the technology in the capitalist sector. They have no other alternative, such as labor markets. Total output and output per worker are then fully determined. Total output produced in the z-subsistence sector is part of national income and its distribution.

Given that the z-subsector output is thus determined independently of the other sectors, the solution found for the omega model will fully apply to the sigma model. This model is very simple and can be represented graphically. Figure 5.1, panels (a) and (b), in Chapter 5, represent the capitalist sector of sigma society as well. The core of the general equilibrium in sigma society can be represented by the labor market equilibrium in terms of real wages. This is represented in Figure 6.1. It is similar to Figure 5.1(c), except that now includes the z-subsistence sector, with marginal productivity of labor equal to the curve $m'n'$ and labor supply equal to $O'Z$. National income is now the sum of three areas: the area under segment HE (capitalist sector), area under segment mf (x-subsistence sector) and area under segment $m'f'$ (z-subsistence sector).

National income (Y) of equilibrium and its distribution between social groups (where x-workers and z-workers are explicitly distinguished by the sub-indices) can be represented by the following system of equations:

$$\begin{aligned}
 Y &= (P + W_x + V_x) + V_z & (6.2) \\
 W_x &= w_x D_{hx} \\
 V_x &= v_x L_x \\
 V_z &= v_z L_z \\
 S_{hx} &= D_{hx} + L_x + U_x \\
 S_{hz} &= L_z
 \end{aligned}$$

The first equation shows that national income is equal to the total output produced in the omega sector (indicated within parenthesis) plus the total output in the z-subsistence sector. As shown in Chapter 5, in the omega sector, total output is equal to output of the capitalist sector and output in the x-subsistence sector; output of the capitalist sector is distributed to capitalists as profits (P) and to x-workers hired in the capitalist sector as wages (W). The third and fourth equations just show that total income of self-employed workers can be decomposed as average income multiplied by the total employment in each subsistence sector. The last two equations show the allocation of labor supply: x-workers are in part employed in the capitalist sector as wage-labor, and the excess labor supply are in part self-employed in the x-subsistence sector and in part unemployed; whereas z-workers are all self-employed in the z-subsistence sector, as they are excluded from the labor market.

In equilibrium, as shown in the omega model, the mean income in the x-subsistence sector must be lower than the wage rate ($w_x > v_x$) because this is a necessary condition for capitalism to function in an overpopulated society. On the other hand, inequality in the endowments of human capital between x-workers and z-workers leads to the following relation: output per worker in the x-subsistence sector is higher than it is in the z-subsistence sector society. Then we get this relation: $w_x > v_x > v_z$.

Again, the subsistence sector plays a crucial role to make viable the functioning of capitalism in an overpopulated society. If the subsistence sectors did not exist, the capitalist sector would not be socially viable in the sigma society. Profits and wages would have to be distributed to the excess supply of workers for their living; moreover, this redistribution would have to reach not only x-workers that are unemployed and self-employed, but to all z-workers as well.

As in the omega model, this sigma model also shows that x-workers who are excluded from the labor market could seek to set up firms or expand the size of their current productive units where they are self-employed. Skills do not constitute a barrier because x-workers are endowed with high-level human capital. They could rent capital to open a firm; but there is no market for capital service. They would have to buy physical capital, which in turn needs financing. But these workers are also excluded from the credit market and the insurance market, making these projects unviable.

Z-workers face even stronger limitations to escape from the relative poverty situation. Renting or buying physical capital is unviable, as they are constraint by their human capital endowments and also by the exclusion from the credit and insurance markets. Accumulation of human capital is also limited and by the same set of constraints.

The theories of exclusion of credit and insurance markets will be presented in Chapter 8 below.

In sum, the general equilibrium shown in Figure 6.1 is indeed an equilibrium situation: no one has both the power and the will to change it. General equilibrium is thus reached and the values of the endogenous variables will be repeated period after period as long as the values of the exogenous variables remain fixed.

6.4 Empirical Predictions

Given the autonomy of the z-subsistence sector in the general equilibrium, the empirical predictions of the omega model established in Chapter 5 will also apply in this sigma model. However, the effects of the exogenous variables upon national income and its distribution will be new due to the presence of z-workers.

In the short run, the relevant exogenous variables are still the international terms of trade and the international interest rate, together with the initial inequality. The effects upon the capitalist sector and the x-subsistence sector will be similar to those we found for the omega model. National income in the sigma model depends positively upon the international terms of trade. The effect of the international interest rate upon national income will be negative.

The effects of the exogenous variables upon income distribution of equilibrium will be a bit more involved. There are five social groups that can be distinguished according to average incomes: capitalists, wage-earners, the self-employed in the x-subsistence sector, the unemployed, and the z-workers. Sigma is a more complex society than epsilon or omega. Just to simplify the analysis, the group of the unemployed can be ignored.

The positive effect of the international terms of trade upon total output in the capitalist sector implies an increase inequality between the capitalist sector and the z-subsistence sector. However, the change in the functional income distribution will be ambiguous within the capitalist sector (as shown in the omega model); thus the global functional inequality will also be ambiguous. Regarding personal income distribution, the result is also ambiguous because both the average real wage rate and the average income in the x-subsistence income will increase, whereas the average income in the z-subsistence sector will remain fixed, and the average profit is ambiguous. Therefore, the effect of the international interest rate upon distribution will also be ambiguous.

The effect of changes in the initial inequality (higher concentration in the ownership of physical capital) is nil upon total output in the capitalist sector and thus in national income. The effect will also be nil upon the functional income distribution. However, personal income distribution will increase because total profits will become more concentrated in fewer capitalists.

The reduced form equations for total output and income inequality in sigma society can then be written as follows:

$$Y^0 = F^\sigma(z^*, r^*, \delta; K_b, S_h) \quad (6.3)$$

$$F_1 < 0, F_2 < 0, F_3 = 0, F_4 > 0, F_5 > 0$$

$$D^0 = G^\sigma(z^*, r^*, \delta; K_b, S_h) \quad (6.4)$$

$$G_3 > 0 \text{ and } G_1 = (?) \text{ otherwise}$$

Equation (6.3) indicates that factor endowments determine the level of national income, with short run variations around that level due to the effect of changes in two exogenous variables: the international terms of trade and the international interest rate. Similarly, equation (6.4) indicates that the initial inequality (which incorporates the effect of changes in factor endowments) determines the level of inequality in the distribution of national income, with short run variations around that level due to changes in the other exogenous variables. The beta propositions are given by the signs of the partial derivatives. (The mathematical proof appears in Appendix A.)

Two additional comments are in order. First, given the assumptions on the asset endowments of z-workers, changes in prices and quantities in the capitalist sector do not affect the z-subsistence sector, directly or indirectly. On the other hand, changes in the z-subsistence sector do not affect the capitalist sector. This sigma model predicts economic dualism between the capitalist sector and the z-subsistence sector. Output growth in the capitalist sector has a “trickle down” effect on the x-subsistence sector, through the labor market, but not on the z-subsistence sector. The x-subsistence sector is residual to the workings of the capitalist sector, but the z-subsistence sector is completely excluded from the capitalist sector.

According to the sigma model, the essential factor that increases average real income of z-workers is the increase in the labor productivity. The model predicts that as long as labor productivity level remains unchanged (curve $m'n'$ in Figure 6.1(c)), the average income of z-workers will also remain unchanged; and so will average income. The situation is different for the x-subsistence sector, in which average income can increase even if labor productivity in the sector remains unchanged. The sigma model predicts that an increase in employment in the capitalist sector will reduce the size of the self-employed in the x-subsistence sector, and thus increase the average productivity of labor in this sector.

Second, what does a change in the initial inequality (δ , the Greek letter delta) imply in sigma society? Redistribution of physical capital endowments from capitalists to workers is one way to reduce δ . Its effect on total output is nil, but it will reduce income inequality because part of profits will now go to workers. Surely, the individual endowments of human capital cannot be redistributed.

Another way to reduce δ in sigma society is to equalize the unequal endowments of political rights in society. This change will have no effects upon private goods; that is, upon prices and quantities in the market system. But it will change the distribution of public goods. So far we have ignored public goods. Assume now that sigma society has two *local* public goods of different quality, instead of a single *universal* public good. As second rate citizens, z-workers are entitled to the low-quality public good. National income, which includes private and public goods, will now show a higher degree of inequality. If z-workers were entitled with the same political rights than the rest of the population, they would have access to the high-quality public goods. Consequently, national income would now show a lower degree of inequality.

Equalization of political entitlements in sigma society will also reduce income inequality through its effect upon the labor productivity in the z-subsistence sector. In the short run, the access of z-workers to existing public goods, or to public goods of higher quality, such as social infrastructure, health services, and judiciary system, will shift upwards the labor productivity curve of the z-subsistence sector (curve $m'n'$ in Figure 6.1); hence, average income will increase. In the long run, the major effect of equalization in political entitlement will operate through the accumulation of human capital, which will shift upward the labor productivity curve and will also generate social mobility as some z-workers will become x-workers (as will be analyzed later on). Thus, a reduction in the initial inequality of assets in sigma society will cause a fall in the degree of income inequality in the short run.

6.5 Empirical Consistency

From a historical perspective, as shown in Chapter 2, the Third World countries can be divided into two groups: those that have a significant colonial legacy and those that do not. While the omega theory intends to explain the latter type of countries, the sigma theory refers to the former type.

Sigma theory, therefore, intends to explain the production and distribution process in those Third World countries that have a significant colonial history. A colonial system established institutions in the colonies, such as the rule of property rights. Another important rule is the political inequality: the rule of the first class citizens (the colonizer) and the second class citizens (the colonized and the slaved brought from Africa). Sigma theory assumes that the inequality in citizenship under capitalism comes from the legacy of colonial institutions. Historians implicitly recognize this legacy by pointing out the nature of the colonial institutions: “Colonial societies were generally characterized by apartheid and segregation and often were based on notions of innate racial inequalities” (Wesseling, 2004, pp. 242-243).

The empirical predictions of the sigma model for the short run can now be confronted against the empirical regularities listed in Chapter 2. The relevant regularities for confrontation include Facts 2, 3, and 4.

The sigma model predicts that the equilibrium conditions of production and distribution are equal to those of the omega society together with an independent z-subsistence sector. The omega model was able to predict Facts 2 and 4; hence the sigma model also predicts these Facts; that is, the sigma model explains why the Third World operates with significant under employment and why changes in nominal variables affect the level of output and wage employment.

Regarding Fact 3, the sigma static model predicts that, in the short run, changes in the exogenous variables will not affect the subsistence sector in which the z-workers are self-employed and thus they will remain as the poorest social group in sigma society. The z-population corresponds to the descendants of the subaltern colonized populations, which include the populations that were subject to slavery traffic and the native populations of the territories that were colonized. Fact 3 says that the average income of the descendants of the subaltern populations is the lowest among social groups in the Third World countries.

The sigma model indeed predicts Fact 3. In sum, the empirical regularities, Facts 2 to 4, do not refute the sigma model.

In the international literature, the observed self-employment in the Third World is usually seen as the *informal sector* or the *shadow economy*. According to this view, this is the result of economic choice made by individuals between the cost and the benefit of legality. The excessive cost of legality over its benefits leads these individuals to act outside the legal system (cf. Schneider & Enste 2000). This illegal, informal sector, or shadow economy could disappear or decrease significantly if the governments were just able to reduce the cost of legality. The hypothesis is that the informal sector is the result of a government failure of over-regulation, which makes costly to have property rights legalized and license to operate firms legally. However, this hypothesis is not a beta proposition derived from a theory (no theoretical construction is available) and, of course, the hypothesis has not been tested statistically.

According to the omega and sigma theories, the observed self-employment in the Third World constitutes the subsistence sectors. They owe their existence to structural factors, such as overpopulation, not to the cost of legality. Just to make sure, costs of legality do exist in the real world; but according to these models, they are not the essential factors explaining this phenomenon. If these costs were eliminated, the size of the subsistence sectors would be reduced, but not by much; the subsistence sectors would still be there. In the omega and sigma models, the costs of legality are ignored. Given the consistency between the empirical regularities and the predictions of the models, the abstract societies omega and sigma are indeed good approximations of the Third World countries.

In the omega society, the subsistence sector played a significant role in general equilibrium: it made possible the functioning of capitalism in an overpopulated society. In the sigma society, there are two subsistence sectors. The first corresponds to skilled workers (as in omega) and the other to unskilled workers. Do z-populations play any role in the functioning of capitalism in the Third World?

Z-populations seem to play the following roles. As part of the peasantry, one role they could play is the supplying of cheap food to the market, as cheap wage-goods. But the proportion of the total food supplies coming from the peasantry is usually small. Another role would be to supply cheap labor, as temporary labor; but again this is a relatively small part in the labor market. The sigma model predicts that whatever those roles may be, they are not significant for the functioning of the capitalist sector. It is *as if* z-workers were indeed the underclass.

There is, however, one aspect in which the z-workers play a significant role in the viability of capitalism in overpopulated societies. They take care of their livelihood to make possible capitalist profits. Suppose that for some reason the lands of the z-peasantry became totally unproductive; the peasantry would now become part of the unemployed workers. Somehow society would have to accommodate this increase in unemployment and possibly a new type of society would appear, with new rules, and new type of general equilibrium. The overpopulated society could not operate with a capitalist sector when the unemployment rate is very high, even if the unemployed is composed mostly of z-workers. Profits and wages are of equilibrium under the condition that z-workers take care of their livelihood.

In sum, sigma theory seems to explain the Third World with strong colonial legacy. The Empirical facts do not refute the model and, therefore, there is no reason to reject this model, and we may then accept sigma theory at this stage of our research.

Up to now, short-run static models of the epsilon, omega, and sigma theories have been presented. The empirical predictions derived from each model have not been refuted by the relevant empirical regularities of the First World and Third World countries listed in Chapter 2. Facts 1 to 4 have thus been explained by these partial theories of capitalism. In order to confront the theories with Facts 5, 6 and 7, dynamic models showing endogenous capital accumulation (physical and human capital) and technological change are needed. Then we will be able to know whether or not the initial inequality of societies will change endogenously in the process of economic growth. On the initial inequality, two questions arise. First, what is the relation between inequality and social order? Second, does inequality play any role in the accumulation of physical and human capital? These questions will be analyzed in the following three chapters. The relationship between inequality and social order will be the theme of the next chapter.

Table 6.1. Social Structure of Sigma Society: Ethnicity, Class, and Citizenship

Ethnic Group	Physical capital	Human capital	Citizenship	Name of Social Group
Blues	K_b	K_{h1}	C_1	A
Purples	0	K_{h1}	C_1	X
Reds	0	K_{h0}	C_0	Z

Symbols:

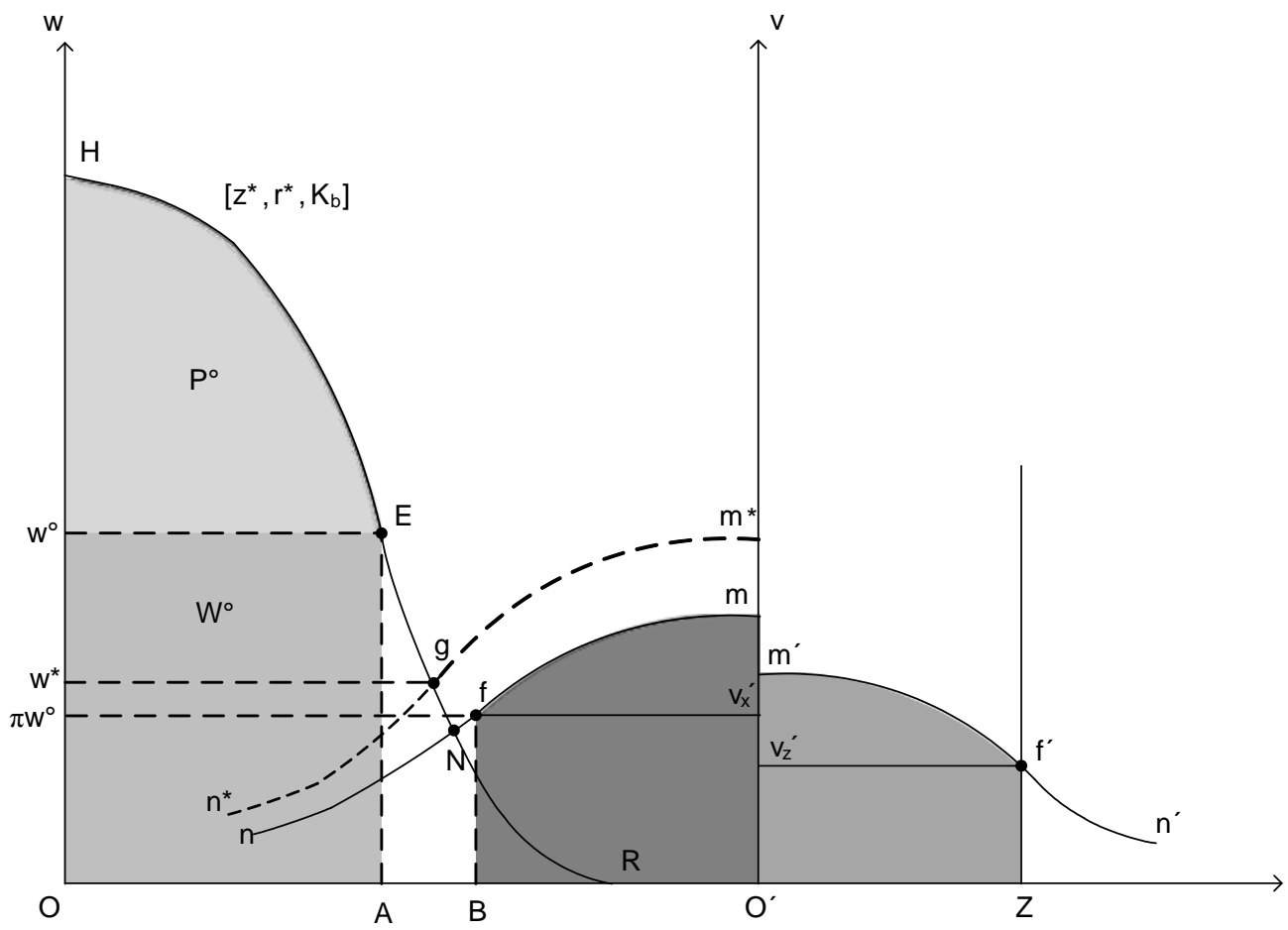
K_{h1} : Human Capital – High Level

K_{h0} : Human Capital – Low Level

C_1 : Citizenship – High Degree

C_0 : Citizenship – Low Degree

Figure 6.1. General Equilibrium in the Sigma Society



PART II

TOWARDS A UNIFIED THEORY OF THE CAPITALIST SYSTEM

CHAPTER 7

INEQUALITY AND SOCIAL DISORDER

According to the epsilon, omega, and sigma models presented so far, both production and distribution in capitalist societies are endogenously determined. Income inequality is thus an outcome of the economic process. An implicit assumption of these models were that capitalist societies can tolerate any degree of inequality that results from the market system; that is, there are no limits to inequality. In these models, there were no rules that set limits to the degree of income inequality. Therefore, once general equilibrium was reached, production and distribution could be repeated period after period under the same rules of the economic game; that is, the general equilibrium implied social order.

Social disorder will be defined as the behavior of people that seeks to break the rules of the institutional context. Could social disorder be generated by excessive income inequality? How can capitalism operate with social disorder? This chapter presents, first, a theory that intends to explain the role of inequality in the generation of social disorder in different capitalist societies; second, it also presents a theory of government behavior towards inequality. Then both theories are integrated into new general equilibrium models of the epsilon, omega, and sigma theories. The empirical predictions are derived and then are confronted against facts.

7.1 Limited Tolerance to Inequality

In order to understand the origin of social disorder, a general theory will be presented in this section. We may call it the *theory of limited tolerance to inequality*. The following alpha proposition is then proposed:

$\alpha(C).(1)$ *Limited Tolerance to Inequality*: Individuals have a sense of justice or fairness with respect to economic inequality, which sets thresholds of tolerance to inequality. Whenever inequality reaches a level beyond their tolerance thresholds, individuals will protest and seek to restore inequality to the tolerable degree.

The symbol “C” after alpha indicates that this theory is “general” in the sense that it is intended to be valid in the three capitalist societies that have been studied, namely, epsilon, omega, and sigma. (The symbol “C” will be applied to other general theories later on.)

Envy and resentment are embedded in this economic logic. In neoclassical theory, these human sentiments fall within the category of externalities. The theory proposed here departs from neoclassical theory in that it assumes the existence of a threshold level of tolerance to inequality. Whenever inequality reaches values that go beyond the threshold, individuals will consider inequality unjust or unfair as judged by themselves. In addition, the theory assumes that people do not just contemplate injustice, but that people react, do something, protest, and resist what they consider to be distributive injustice. The existence

of threshold of tolerance is critical in the theory; if such threshold did not exist, any degree of inequality would be socially tolerable and thus extreme inequality would not generate social disorder.

A particular model of the limited tolerance theory is now developed. Consider the following utility function for individual j :

$$U_j = U_j(Y_j, Y_j/Y_k) \quad (7.1)$$

Such that

$$U_j \geq 0 \text{ if } (Y_j, R_j) \geq (Y_j^*, R_j^*)$$

$$U_j < 0 \text{ if } (Y_j, R_j) < (Y_j^*, R_j^*)$$

Here Y_j and $R_j = Y_j/Y_k$ are the individual's absolute and relative incomes, where k is his group of reference. The values marked with asterisk represent the individual's threshold values of tolerance to absolute and relative incomes. The model also assumes that the threshold of absolute income is the current income. Therefore, only the threshold value of the relative income is unobservable, which means they are *exogenous elements*, not exogenous variables; so, they must be assumed to be fixed.

This utility function assumes that the individual considers both types of incomes as goods and thus more quantities of each are preferred to less. The model assumes that, in addition to commodities, the individual seeks to consume social status. Beyond the threshold values, the individual is willing to give up a part of his absolute income in exchange for a higher relative income and vice versa. Below the threshold values, any pair of absolute and relative income has a negative value of utility.

The utility function thus assumes that individuals have a sense of fairness when it comes to absolute and relative incomes. The negative value of utility indicates that this situation is non-tolerable and that the individual will react and protest. It indicates not only a subjective pain caused by this situation (just a case of externality in his utility function), but assumes that the person will take actions that seek to restore inequality to the tolerable range. The actions will include breaking the institutional rules, for the individual will consider the excessive inequality the result of an unfair institutional framework. Private property rights and democracy rules will be challenged. Those actions, in the aggregate, lead to *social disorder*.

This logic is also consistent with Maslow's theory of hierarchy of human needs (Maslow, 1970). It reflects that people's primary needs are associated with physical survival as well as social survival; that is, the utility function assumes that people do not passively accept extreme situations of absolute poverty or relative poverty (inequality).

According to this theory, there is a subjective and objective part of too much inequality: individuals that experience the injustice will develop sentiments of resentment that will originate destructive impulses and ultimately actions of protest. On the side of winners, the fears of protests will also emerge and lack of trust in the human relations within society will also develop. Social environment becomes more insecure and full of fears under too much inequality.

The assumption of self-interest motivation implies that individuals are selfish, greedy, and egotistic. Individuals prefer more income to less. Therefore, income

distribution in a capitalist society is the major source of social conflict. Distributive injustice will then lead to severe social conflict and then to social disorder.

The set of income distributions that are tolerated by all social groups will be called the *region of social tolerance to inequality*. This is a strict subset of all possible distributions of income; that is, the region consists of various distributions, not a unique distribution. This model is illustrated in Figure 7.1. Society is composed of two social groups, A and B. The region of social tolerance is D^* . Perfect equality (point E) does not belong to the region, as the wealthy (group B) will not tolerate this situation. Situation C will not be tolerated by group A. Figure 7.1 assumes that tolerance to inequality is determined in terms of proportional differences in income. (We could also assume tolerance in terms of absolute differences in income, which would define D^* by parallel lines to the 45° line.)

The region of social tolerance does not imply that individuals are indifferent among every possible income distribution within the region. This is still an area of social conflict about income inequality, in the sense that social groups would prefer a higher share of total income; but region D^* do not lead to social disorder, as all social groups would accept this distribution. However, this region could not be called “equilibrium inequality” because there is no mechanism by which income inequality that is outside the region would return to the region spontaneously, as will be shown below.

In the models of epsilon, omega, and sigma theories presented so far, income inequality fell implicitly into the region of social tolerance. But it need not. There is nothing in those models that assures that income inequality will fall necessarily within the region of tolerance. Moreover, if income distribution lies outside the region of tolerance, there is no market mechanism that can return it to the tolerance region.

Assume that individuals have different thresholds of tolerance to inequality. For a given degree of inequality, there will be a group of people that tolerates it and another that does not. If the degree of inequality increases further, the group that did not tolerate the initial inequality will certainly not tolerate this higher inequality either; but from the group that did tolerate the initial inequality, there will be a subgroup that will not tolerate this higher inequality. Therefore, there will be more people that do not tolerate the higher degree of inequality.

The empirical prediction of the model is that in the aggregate there exists a positive relation between income inequality and social disorder. The empirical prediction between social disorder (SD) and the degree of inequality (D) can then be written as follows:

$$SD = f(D), f' > 0 \quad (7.2)$$

A comparison of this theoretical model with other theories on inequality tolerance is in order. First, consider the theory of justice developed by the philosopher John Rawls (1971). His normative theory can be transformed into a positive economic theory as follows. Consider a society in which individuals face the same context, in which everybody may go through the same risk of losses or the same windfall gains; in such society, the distributive rule of consensus will be:

1. Everybody is subject to the same risk of having no income; then people will agree on setting rules for a guaranteed minimum income for all, as a protection against that risk.
2. Everybody is subject to the possibility of extraordinary gains; then people will agree on setting redistribution rules so as to share part of that gain.

In such context, any individual led by his self-interest would independently design this distributive rule. It is an impersonal rule; it is symmetric; it is fair; thus, it is just. It is acceptable as part of the social contract.

Rawls' theory implies a region of social tolerance that is symmetric with respect to the 45° line in Figure 7.1. Hence, a beta proposition of Rawls' theory is that perfect equality would belong to the region of tolerance; that is, perfect equality in society will lead to social order. This is a different prediction from what emerges from the theory of limited tolerance to inequality, in which the wealthy would not tolerate perfect equality.

Implicitly, the abstract world of Rawls is one in which there are no social classes. In this society, the probability of reaching any position in the distribution of income is the same for every one; people play the economic game with unloaded dice. This is his assumption of people acting upon a veil of ignorance about the future. It is a rule for very flexible and open societies, where there are no institutional barriers that limit social mobility (where the Spanish expression *la tortilla puede darse la vuelta* would apply). In a society of class and citizenship differences, this could hardly be the case. The opportunities for earning incomes are different for different people. This is due to the initial inequality, that is, the inequality in the individual endowment of economic and political assets.

Another important author on this topic is the economist Albert Hirschman (1973). He proposed the theory that any inequality is tolerated if it is expected to be temporal. In the initial stages of economic growth, when inequality is likely to increase, society's tolerance to such disparities will be substantial if people's expectations are that those disparities will only be temporary. If those expectations are not met, there will be social disorder. Hirschman uses as a metaphor a tunnel in which there is traffic blockage. When cars start moving in one of the lanes only, people in the other lane will accept this disparity if they expect that shortly they will be moving as well. If this does not occur, social protest will emerge.

Hirschman's theory can be interpreted as the transitional dynamics of moving from one situation of inequality to another of higher degree of inequality. Over time, social disorder will be higher because more people will not see their expectations fulfilled. It is clear that Hirschman's theory is complementary to the limited tolerance theory, for it adds an empirical prediction about the time path of social disorder as inequality increases.

7.2 Government Behavior towards Inequality

What would be the government behavior towards inequality and social disorder?

Regarding equilibrium income distribution in epsilon, omega, and sigma models, the implicit assumption so far has been that the net tax incidence is neutral on inequality; that is, given the constraints faced by governments, fiscal policy on expenditure and taxes leaves unchanged the inequality generated in the market system. Thus, there was no need to introduce government behavior into the general equilibrium models and it has been safely ignored. In addition, it has implicitly been assumed that the equilibrium income inequality always falls within the social tolerance region.

Social order may be considered a public good. It is a good that is non-rival and non-excludable. Nobody can be excluded from consuming social order. The provision of public goods usually falls in the domain of the state because of the market failure to produce them. Firms have the incentive to produce only goods that are privately consumed, which can then be sold in the market, and which can thus generate profits. (No one can buy some units of social order together with some kilos of potatoes in a super market). Public goods play an important role in society. The presence of public goods helps to construct a human society; if it were not for public goods, a society would just be a sum of individuals. In capitalist societies, governments are in charge of the provision of public goods, such as money, infrastructure capital, and social order.

Standard economics usually assumes that government behavior is exogenously determined in the economic process. Governments are even told what to do. Consider the fact that every economic study concludes with a section or chapter about policy recommendations, in which the study tells governments what they should do in order to solve a particular problem. This literature assumes, implicitly, that governments do not pursue interests of their own.

Another view within standard economics comes from public choice theory. The assumption here is that people when in government act guided by the motivation of self-interest, as in any other activity. Politicians may also have other motivations, such as political or ideological, but these are not considered the essential factors underlying their behavior. Politicians seek political power and incomes that go with running the state apparatus; thus, their social function is accomplished as a by-product of their selfish motivation (cf. Downs, 1957; Orchard & Stretton, 1997; Persson & Tabellini, 2000).

In this chapter, the assumption that politicians act guided by the motivation of self-interest will be adopted; however, others assumptions will be added in order to construct a theory of government behavior. The following alpha proposition is intended to be valid in every type of capitalist society:

$\alpha(C).(2)$ *Rationality of Politicians*: Politicians seek two objectives hierarchically ordered: firstly political power and only then incomes, subject to pressure group demands, the fiscal budget constraint, and institutional constraints.

This theory assumes the existence of a political class. As in the case of capitalists, politicians firstly seek to remain in the privileged position of membership of a social class: the political class. The objective of income maximization is subject to the assurance of the first objective. There is no substitution between incomes and political power.

According to this theory, governments act guided by these selfish motivations; they are social actors, as capitalists or workers, and as such they interact with the other social actors; they are not above them. Given the values of the exogenous variables, governments

will choose the values of the endogenous variables according to their motivations. Changes in the exogenous variables will modify their choices.

This theory is intended to be valid, in principle, in the three types of capitalist societies. The government rationality is a general assumption, but the constraints will be different in each society, as will be shown below. The theory is also intended to be valid in all forms of democracy: representative or participatory. Markets and democracy are considered the fundamental institutions of capitalism; however, as in the case of markets, in which the development of markets is endogenous, it will be shown here that the degree of democracy (going from weak representative democracy to strong participatory democracy) is also endogenous.

To put this theory of government behavior into the falsification process, auxiliary assumptions are now introduced to construct a particular model of the theory. First, the model will assume the representative democracy as the political system in all types of capitalist societies. Second, governments will seek to maximize votes (or popularity, which is a form of voting) subject to the public budget constraint, institutional constraints, and the demands of the pressure groups. The components of the public budget (expenditure, taxes, debt) constitute the endogenous variables. Third, votes depend on the performance of the aggregate variables of the economy, such as output, employment, income inequality, and inflation rates.

Given the exogenous variables, governments will seek to reach their best positions, their equilibrium positions, regarding the objective of maximization of votes. This objective will lead to the conditions of equilibrium that the governments will seek. The following government behavior is implied by the model in the search of the equilibrium position:

1. Governments will seek to allocate the public budget to discretionary expenditures (with which to buy votes) rather than to mandatory expenditures (on securing citizens' access to their rights, which do not buy votes).
2. Governments will seek to finance fiscal expenditure with more debt rather than more taxes, as the latter will affect negatively next elections, whereas the former will not.
3. Governments will seek to use fiscal policy according to the political cycle; that is, governments will seek to increase public expenditure before elections and reduce it after elections. Fiscal policy is dependent on the political cycle.
4. Governments will seek to allocate discretionary public expenditure by sectors and regions according to their political profitability; that is, expenditure will be higher in the most densely populated areas (urban rather than rural areas) and in the most noticeable projects (school buildings rather than education reform to increase its quality).

These propositions just refer to the equilibrium situations that governments will seek given the values of the exogenous variables. These equilibrium conditions are observable and thus constitute empirical predictions of the model as well. The model predicts a myopic behavior of governments: they will seek to give priority to the short term effects of public policies on buying votes rather than to the long term effects. Buying votes have priority because this is the way to assure the priority objective of political class membership.

These empirical predictions of the model have not been studied in the international empirical literature. Prediction (3) was corroborated in a study about the U.S. government behavior, in which it was shown that government social expenditure indeed increases before elections and falls after elections (Rogoff, 1990).

How does the government behave toward inequality in each type of capitalist society? American economist Arthur Okun (1975) pointed out that democratic capitalism operates with a double standard: the political system preaches equality, but the market system operates with inequality. In order to make the system viable, Okun assumes that society seeks to establish some rules that set limits to inequality. The state establishes rights aimed at that objective. Some goods and services are taken out of the market place and distributed to the population as public goods. In this way, money cannot buy everything in society. One could consider this an institutional rule for the functioning of society. This rule would be part of the social contract.

The question is whether social actors have the required power and incentives to establish this type of social contract. Where do the political demands and the provision of rights come from? Okun does not solve this question. Implicit in Okun's theory is that the society he is referring to is an epsilon society or omega society, in which people are citizens of the same class and can, therefore, demand universal rights. The situation would be different in a sigma society, in which second rate citizens are the most in need of rights, but they have no voice to demand them precisely because they are excluded from rights. They lack the rights to have rights.

The government behavior model proposed here will assume that government behavior will be different in different types of capitalist societies. The model therefore predicts that mandatory expenditure (allocated to the provision of rights in the form of public goods) as proportion of total government budget will be higher in epsilon society than in sigma society; moreover, the provision of public goods as part of total output will also be higher in the former. This prediction seems to be consistent with the facts we observe in the First World and the Third World.

What will governments do in sigma societies, where income inequality tends to be high? In order to assure social order, governments may use a mix of repression and redistribution measures. The former has immediate results to restore social order, whereas the latter will show results in longer periods, maybe after next elections. Besides, the constraints put by the wealthy, a powerful pressure group, will go against redistribution. Governments will then choose more repression and less redistribution to restore social order.

This equilibrium condition in the behavior of governments may seem counter intuitive. One may expect that in a democratic capitalism the majority or the median voter rule would endogenously reduce inequality to more socially tolerable level. A democratic decision is based on the majority rule, which implies that the opinion of the median voter (which is placed at the center of all positions) will decide. In the case of income redistribution, similarly, the opinion of the median voter will decide. But does the median voter belong to the poor or rich groups? If he belongs to the poor group, then the redistribution policy will win.

If the distribution of income is a normal distribution—asymmetric distribution or bell shape distribution—the mean income will divide the population into two equal parts

(50% below the average and 50% above the average); being the distribution symmetric, for each rich individual (with income above the average) there will be a corresponding poor individual. Well this is the same definition of the median value (50% is above the median and 50% is below the median), as the mean and median are equal. However, when we talk about the income inequality in the real world, we refer to the case of income distribution that is not symmetrical: the mean income will divide the population into two unequal parts (say 70% below the average and 30% above) because for each rich individual (with income above the average) there will correspond many poor individuals. Therefore, the median income (dividing the population 50% below and 50% above) will be smaller than the mean income, which implies that the median voter will belong to the poor group!

Democracy should then produce redistribution policies endogenously and income inequality will be a self-regulated variable. Not by the market system, but by the democratic system. If inequality rises, public policies chosen democratically will bring it back to a more socially tolerable value and social disorder could hardly have any significance. However, this is not what we observe in the real world, neither in the First World nor in the Third World.

Why does the median voter principle fail? There are several reasons. Under representative democracy, politicians are given power. Voters do not choose policies, their representatives do; moreover, the representatives are not a random sample of their voters, as doctors are not of their patients (Banfield 1958). Some political scientists have proposed the hypothesis that capitalist democracy takes the form of plutocracy: it is the government by the wealthy and for the wealthy (Fukuyama 2011). Voters and representatives face the *principal-agent problem*. This problem appears whenever the objectives of the agent (the government in this case) are inconsistent with those of the principal (the voters).

In sum, the political class model predicts that governments behave differently in epsilon, omega, and sigma societies. In epsilon compare to epsilon, there are few rules to set limits to inequality that the government must implement; discretionary expenditure is therefore relatively more available. Discretionary government expenditure is the instrument to buy votes and it is relatively more significant in sigma society. In general, income inequality is not a self-regulated variable in a capitalist society: too much income inequality will not be reduced endogenously by public policies. Governments have no incentives to bring it to the socially tolerable region. This problem is more significant in sigma society.

7.3 Social Order as Production Factor

A public good usually refers to consumption goods. However, a public good may also be defined by the characteristic of its production; it is a good that is produced collectively but can then be exchanged as private good. This is the case of public education. The good human capital that is produced publicly can be exchanged in labor markets by individuals. This is different from a bridge, which is produced collectively and used also collectively, as most pure public goods are. Social order is a public good that may be defined by using both criteria. It is produced collectively and it is consumed collectively.

A new assumption is now introduced about the production process. Apart from machines and workers, some public goods also constitute factors of production, such as human capital, infrastructure capital, and social order.

Aggregating over all capitalist firms of the economy, consider the production function of the following form:

$$Q_b = F(L, K_b, K_h, K_i, O), F_i > 0 \quad (7.3)$$

This equation says that the quantity of good B produced (Q_b) depends upon the quantity of workers hired in the labor market (L), the endowment of capital stock of the firms (K_b), and the stock of public goods, which is supplied by the state. Public goods include the stock of human capital (K_h), the stock of physical infrastructure (K_i), and the degree of social order (O). In addition, the quantity of labor inputs (D_h) is decomposed into human capital (K_h) and quantity of workers employed (L).

How does social order affect the production process? At the capitalist firm level, social disorder has a negative effect on the production process. With social disorder we cannot say that “the production process can be repeated period after period as long as the exogenous variables remain fixed.” Some losses in the flow of output or in the stock of capital will occur due to social disorder. Social disorder affects the normal production process through periods of interruption due to workers strikes and social revolts, which implies shocks that reduce the flow of output compared to periods of social disorder absence. Social disorder implies no respect for the rule of private property rights.

Social disorder thus implies that workers will take actions to reduce inequality through forced private redistribution of income or assets. This forced private redistribution is equivalent to a lump sum tax levied upon the capitalists. Production is now subject to risk and variability. Due to these possible shocks, total output and labor productivity in the firms are now measured in terms of their mean values and their risks. In sum, social disorder implies a fall in total profits.

Under this environment of social disorder, what will capitalists do? Because capitalists seek to maintain their privilege as members of the capitalist class, they will react to the situation by acquiring more inputs to protect their property rights. Some quantities of inputs destined to protect private property (such as fences and security systems) will be added to the production process, even though they are not necessary from the technological point of view. Additional workers will be hired as *protective employment*. Insurance policies will also be bought in insurance markets. Because capitalists also seek to maximize profits, firms will be willing to increase their production costs in the form of protective fixed costs if the resulting total profits will be higher compared to doing nothing.

Assuming that the effect of social disorder on capitalists (including their reactions mentioned here), takes the form of additional fixed costs, total employment will increase due to the use of protective employment, although productive employment will not change. Fixed costs associated to protecting the production process do not affect the marginal productivity of labor, which is critical to determine the productive employment level. The consequence at the firm level is that, on average, the same quantity of output will be produced using higher quantities of labor and physical capital inputs. But the additional fixed costs will reduce profits. Output per worker will fall.

What is behind social disorder? It is income inequality, as suggested by the model of the theory of limited tolerance to inequality. The degree of income inequality in society is an exogenous variable for individual firms. The consequence of social disorder at the

firm level is that the same quantity of output will be produced using higher quantities of labor and physical capital. Output per worker will fall. Therefore, income inequality affects negatively labor productivity and total profits.

7.4 General Equilibrium with Social Disorder

At the aggregate level, the same level of output in the capitalist sector will be produced utilizing more quantities of labor and capital inputs than are technologically required. Output per worker will fall. The excessive inequality, and the resulting social disorder, reduces the efficiency of the economy. The higher the degree of inequality is, the higher the losses in the aggregate economic efficiency will be, and also the lower the profit levels will be.

From the theory of limited tolerance to inequality, it was derived the empirical prediction that the degree of social order depends on the degree of income inequality. Moreover, as it was shown in the previous chapters, in all kinds of capitalism the theories predict that the degree of income inequality (D) depends upon the initial inequality in asset endowments of society (δ). Hence, making the necessary substitutions, the aggregate production function in the capitalist sector becomes

$$\begin{aligned} Q_b &= F(L, K_b, K_h, K_i, \delta), \quad F_5 < 0, \text{ and } F_i > 0 \text{ otherwise} \\ Q_b/L &= f(k, \delta), \quad f_1 > 0, f_{11} < 0, f_2 < 0 \end{aligned} \quad (7.4)$$

The first equation shows that, given the quantity of workers and the stock of capital in its different forms, total output obtained by firms will depend upon the degree of initial inequality.

The second equation shows the average productivity of labor, which is derived from the first equation under the assumption of constant returns to scale technology and aggregating all forms of capital into a single composite factor “capital” K . (We can invoke the Hicksian composite good theorem: a group of goods can be treated as one good if the relative prices are held constant.) Given the degree of the initial inequality δ , output per worker depends positively upon the capital-labor ratio ($k=K/L$); given the capital-labor ratio, output per worker depends negatively upon the initial inequality. Other things being equal, highly unequal societies will have a productive system that is less efficient compared to more equal societies.

In the formulation of the production function presented here, where output also depends on the degree of inequality, a new category of factor intensity appears: a particular industry can be considered more (or less) “equality-intensive” compared to other industries. This intensity depends upon the technological requirement for social order in the production process of this industry. Industries that use complex systems of production, such as strong inter-linkages with other industries, will tend to be equality-intensive industries (e.g., manufactures), compared to other industries that can be produced in isolation or in enclaves. In the latter case, the requirement for social order will be less significant and the industry is less equality-intensive (e.g., oil fields, mine centers, tourist centers, maquila-zones).

As pointed out above, in the models of epsilon, omega, and sigma theories presented in Chapters 4 to 6, there was an implicit assumption: the income inequality of equilibrium

always fell in the region of social tolerance; that is, general equilibrium was with social order. Thus, in the static system, the economic process was repeated period after period, with the same output and the same degree of income inequality, as long as the exogenous variables remained fixed.

In order to introduce social disorder in the analysis, we need to see whether general equilibrium with social disorder can be generated. It is now time to integrate the findings of this chapter into the construction of new models. In this chapter, we have developed three structural relations in the workings of each type of capitalist society:

1. Total profits and labor productivity (including productive and protective labor) depend inversely upon the degree of social disorder.
2. The degree of social disorder depends directly upon the degree of income inequality.
3. Therefore, total profits and labor productivity depend inversely upon the degree of income inequality.

These relationships incorporate the prediction derived from the model of government behavior, which says that governments lack the power and the incentives to reduce the income inequality generated by the market system.

Consider a sigma society. Suppose that the equilibrium degree of inequality found is not socially tolerable; then it can be seen now as an intermediary equilibrium situation, which needs to go through another round of interactions to reach the final general equilibrium. It should be noted that the income inequality found in the intermediary equilibrium is the one that results from the given exogenous variables and the market and democratic institutional rules. The degree of income inequality therefore includes the net income transfers (transfers minus taxes) to social groups made by the government. The implicit assumption so far has been that the net transfer is neutral on inequality; that is, given the constraints faced by governments, the fiscal policy on expenditure and taxes leaves unchanged the degree of inequality generated in the market system.

Given that the general equilibrium generates a degree of income inequality that is not socially tolerable, in the first round social disorder and forced private redistribution actions would set in and profits would fall, as indicate above. In the second round, governments would react with repression and capitalists with measures to protect their private property rights, which would increase fixed costs of production. Then social disorder would be controlled to some extent, and now profits would not fall as much. Assume that these interactions converge to a final equilibrium, in which total profits are not as low as it would have been in the first round, but smaller than they were in the initial situation.

The same can be said about income inequality and social disorder. In the new equilibrium situation, income inequality is not as high as it was at the initial equilibrium situation. Some forced private redistribution has taken place, although not in the same magnitude that was intended in the first round. Social disorder will not be as strong as it was at the initial situation.

So the equilibrium values of equilibrium output and profits will correspond to their mean values with a variability given by the likelihood of shocks of forced private redistribution carried out by workers. Two types of employment will prevail: productive employment (technologically required) and protective employment (as inputs to reduce the shocks of forced private redistribution, not technologically required). The security industry will expand. These are the main feedback effects of too much income inequality into the economic process.

The final general equilibrium is therefore *equilibrium with social disorder*. Given the exogenous variables, production and distribution of equilibrium are determined. But now the social environment is one of fear, distrust, and uncertainty due to possibility of shocks of private redistribution that the capitalist class will confront. Because there is no social actor that has the power and the will to change this situation, general equilibrium with social disorder is an equilibrium situation. The equilibrium is static and risky: the mean values and the variability (degree of risk) of the endogenous variables will be repeated period after period as long as the values of the exogenous variables remain fixed. This conclusion applies to epsilon and sigma societies as well.

Figure 7.2 shows the general equilibrium with social disorder in any type of capitalist society. Consider the capitalist sector only. The labor demand curve is HR. The equilibrium situation in the capitalist sector is given by point E, at which the income distribution is socially non-tolerable. The social disorder generated leads to reactions, which are summarized in the hiring of additional labor as protection of property rights. The new equilibrium is at point F, with OB employment, of which AB is protective labor. (Just for simplicity, the market wage rate is maintained unchanged.) The total cost of this additional labor implies a fall in profits, which are indicated by the equality of the two shaded areas. It is as if firms paid for insurance to protect their private property. Therefore, shocks on incomes and property will still occur, but the damage to firms will be smaller. Profits at point F will be smaller than at point E, with a given variance, because of the private redistribution shocks.

In sum, general equilibrium with social disorder does not imply lack of equilibrium; it is a particular form of solution of production and distribution. General equilibrium with social disorder is inefficient, as labor productivity is below its potential and society is a high-risk society, which diminishes the quality of life of people living in this society. It is a second-best solution because social disorder does not challenge the capitalist system, but it is a way to make the system workable. This is not a theory of social revolution, but of social disorder. The nature of general equilibrium with social disorder is similar to other types of awkward general equilibrium of capitalism that have been shown in this study, such as the general equilibrium with unemployment in the epsilon society.

7.5 Empirical Predictions

Under the model of general equilibrium with social disorder, the effects of changes in the exogenous variables on the endogenous variables will shift the equilibrium at point F (Figure 7.2) to another equilibrium situation with protective employment. Because capitalist societies differ in their initial inequality, which determines the degree of income inequality, which in turn determines the degree of social disorder, social disorder will be higher in the more unequal societies. To be sure, equilibrium with the highest degree of social disorder

will be the characteristic of sigma society, while epsilon society will show the lowest degree, and omega will lie in between.

In the particular case of sigma society, equilibrium in production and distribution implies that this society will operate with a large sector of illegal activities, with violence, corruption, excess bureaucracy, unstable political system, fragile and weak institutions, and many other social diseases. Epsilon society will show lower degrees of these social diseases, while those of omega will lie in between. These are the costs of excessive inequality.

Capitalist societies will operate with different degrees of income inequality and social disorder. Income inequality is not a self-regulated variable. Across capitalist societies, across international markets, there is the law of one price (as shown above with the international terms of trade); however, there is no such thing as “the law of one degree of inequality.”

A distinction needs to be made now about the income distribution process in each type of democratic capitalist society. The income distribution that emerges from the market and democratic institutional rules may be called the *primary distribution*. If this inequality falls within the social tolerance region, it will be the end of the income distribution process; if it does not, and as a response to social disorder, public policies may be applied to redistribute income in the form of lump sum transfers. This may be called the *secondary distribution*. If the new inequality falls within the tolerance region, it will be the end of the distribution process. If it does not, private forced redistribution will take place and intend to modify the secondary distribution. This may be called the *tertiary distribution*, and that will be the final static general equilibrium with uncertainty. In Figure 7.2, the primary distribution is represented at point E and the tertiary distribution at point F.

The partial models predict that primary and secondary distributions are similar in sigma society due to the lack of economic rights in society. In epsilon society, the secondary distribution will be more equal than the primary distribution precisely due to the existence of economic rights, such as unemployment insurance. The case of omega lies in between.

Empirical data on income inequality usually refers to primary distribution, although sometimes data are available for secondary distribution; however, data on tertiary distribution is rarely available. The assumption to be made at this point is that the second order and third order distributions will vary around the primary distribution in each type of society.

The theory of limited tolerance to inequality and the theory of government behavior have been introduced into general equilibrium models of the epsilon, omega, and sigma societies. General equilibrium conditions have then been established. The general equilibrium solutions that were presented in Chapters 4 to 6 have now been extended to include the cases of social disorder. The conclusion is that general equilibrium with social disorder is an outcome of the economic process; thus, in the economic process the values of the endogenous variables will be repeated period after period as long as the exogenous variables remain fixed.

In order to derive beta propositions from these general equilibrium models, the condition that the equilibrium is stable in each case must be established. Labor and money

markets still constitute the core of the general equilibrium in each model. In terms of equilibrium conditions, the only difference between the general equilibrium with social order and the general equilibrium with social disorder lies in the composition of employment. In the former case all employment is productive employment, where marginal productivity of labor is equal to the market real wage rate; by contrast, in the latter case, total employment includes productive employment plus protective employment. Because protective employment is a fixed amount of employment (a fixed cost for the firm), the stability conditions are still satisfied because they refer to marginal values.

Comparative statics may therefore be applied to the general equilibrium with social disorder in each model of society. In the short run, the effects of changes in the exogenous variables on the endogenous variables will qualitatively be exactly equal to those found in the case of general equilibrium with social order. The reason is, again, that the productive labor side of the labor market is the relevant equilibrium condition. For instance, if the international terms of trade increases, the productive labor demand for labor will be shifted outwards; hence, the new equilibrium will imply both higher real wage rate and employment level. To this employment level, a fixed amount of employment for protective purpose will be added, which will imply a new equilibrium with even higher values of both endogenous variables. The direction of the effects (the signs) will not change. So the beta propositions for the short run will be similar to those obtained in the case of general equilibrium with social order, except for the effect of the exogenous variable initial inequality.

The national income (Y) of equilibrium for each society can be presented as the following system of equations:

$$Y^0 = F^j(z^*, r^*, \delta_j; K_{bj}, K_{hj}, K_{ij}, S_{hj}), j=\varepsilon, \omega, \sigma \quad (7.5)$$

$$F_2 < 0, F_3 < 0, F_7 \geq 0, \text{ and } F_i > 0 \text{ otherwise}$$

These are the reduced form equations, one equation for each type of society, as indicated by the superscript j , which indicates not only differences in the function, but also in the range of values of the exogenous variables. Two additional forms of capital have been added here: human capital K_h and infrastructure capital K_i . Factor endowments determine the *level* of national income, with short run variations around this level, given by the effect of changes in the other exogenous variables and by private redistributive shocks.

The system (7.5) represents the case of general equilibrium with social disorder, in which the only difference with the previous cases of general equilibrium with social order is the effect of the initial inequality (δ). This effect was nil in the previous cases, but it is negative now. Instead of considering only the two situations studied in this chapter, either with social order or social disorder, equation (7.5) considers a continuous degree of social disorder. This will be the case in a society composed of many individuals, or several social groups, with different thresholds of tolerance to inequality. Therefore, the higher the initial inequality in the distribution of assets is, the higher the income inequality, then the higher the degree of social disorder, and then the lower the output level will be. This is the economic cost that society pays for operating with a high degree of initial inequality.

It was shown in Chapters 4, 5, and 6 that, in each society, the only exogenous variables that changed the level of income inequality were the exogenous variable initial inequality. This result can be presented in the following reduce form equation:

$$D^0 = G^j(z^*, r^*, \delta_j; K_{bj}, K_{hj}, K_{ij}, S_{hj}), j = \varepsilon, \omega, \sigma \quad (7.6)$$

$$G_3 > 0, G_1 = ? \text{ otherwise}$$

In each society, the level of income inequality is determined by the initial inequality, with short run variations around this level, given by the effect of changes in the other exogenous variables and by private redistributive shocks. Income inequality D in equation (7.6) refers, in principle, to primary distribution; however, it applies to any of the three distributions defined above. The assumption is that the second order and third order distributions will vary around the level given by the primary distribution.

On the relations between inequality and social disorder, four beta propositions can be derived from the theory of limited tolerance to inequality. They are:

1. There exists a positive relation between income inequality (D) and social disorder (SD) in each capitalist society. This relation comes from equation (7.2).
2. Between capitalist societies, the degree of social disorder will be higher in sigma than in epsilon. This is the logical consequence of the assumption that the initial inequality determines the norms or institutions of income redistribution in society; that is, because epsilon is socially homogeneous society, there exist norms to set limits to inequality as part of the social norms, whereas this is not the case in sigma. Therefore, the range of variations of the degree of inequality will be different: higher range for sigma compared to epsilon and sigma. Hence integrating equations (7.2) and (7.6), we get

$$SD_j = F_j(D_j) = F_j(G_j(\delta_j)) = \Phi_j(\delta_j), \Phi_j' > 0, j = \varepsilon, \sigma \quad (7.7)$$

$$SD = \Phi(\delta), \Phi' > 0, \text{ for } \delta_\sigma > \delta_\varepsilon \quad (7.8)$$

In sum, on the relations between inequality and social disorder, the model of general equilibrium with social disorder predicts that across capitalist societies, societies with higher initial inequality (sigma type) will have higher degrees of income inequality and social disorder than societies with lower degree of initial inequality (epsilon type). Omega type societies will lie in between.

Equations (7.7) and (7.8) are represented in Figure 7.3, where the degree of primary income inequality D is measured on the horizontal axis and the degree of social disorder SD is measured on the vertical axis. The curve AB corresponds to epsilon, whereas the curve EF to sigma. The figure shows that sigma society would show the highest level of social disorder. This is due to differences in the initial inequalities, which determine social norms, including the workings of the democratic system; that is, less social disorder would be needed to reduce income inequality in epsilon because of the existence of economic rights to redistribute the primary income distribution. In sigma society, by contrast, forced private redistribution and the consequent social disorder is the only mechanism that workers can use to redistribute the primary income inequality.

The consequence is that epsilon society operates in the region of low inequality and low social disorder, point M in Figure 7.3; by contrast, sigma society operates in the region of high inequality and high social disorder, point N. (For simplicity, let the functioning of omega society be similar to that of epsilon because it is also a socially homogeneous society.) The relationship across societies is represented by the curve R, a curve that goes through points M and N. This relationship is empirically observable and falsifiable; it is a beta proposition.

It should be noted that the prediction that more unequal societies show more social disorder leaves open the question of whether social disorder takes the form of individual or collective violence. The theory does not necessarily predict collective violence because this would require collective action, which depends on other factors to materialize (cf. Olson, 1965).

Another prediction on international trade relations is worth mentioning. According to the model developed here, output depends not only upon machines and men, but also upon the quality of society, upon the degree of inequality. Therefore, a new category of factor intensity, different from capital-intensive or labor-intensive industries, will appear: a particular industry can be considered more (or less) “equality-intensive” compared to other industries. This intensity depends upon the technological requirement for social order in the production process of the industry.

Industries that use complex systems of production, such as strong inter-linkages with other industries, will tend to be equality-intensive industries (e.g., manufactures), compared to other industries that can be produced in isolation or in enclaves. In the former case, social disorder will be very costly to the industry productivity; in the latter case, the requirement for social order is less costly and the industry will be less equality-intensive (e.g., oil fields, mine centers, tourist centers, and maquila zones).

The standard theory of comparative advantage in international trade intends to explain trade specialization by connecting factor intensity of industries with factor endowments of capitalist countries. This theory predicts that relatively labor abundant countries will specialize in the production of labor-intensive goods. Introducing the new category of factor intensity, a new prediction on the determinants of international trade can be stated as follows: more equal countries (epsilon societies) will specialize in the production of equality-intensive goods (manufactures); alternatively, more unequal countries (sigma economies) will specialize in the production of less equality-intensive goods (minerals). The initial degree of inequality can therefore be seen as a significant initial condition of the society, as important as the more conventional factor endowments, and it may be responsible for the failure of factor endowments theory to explain the patterns of international trade. (More on trade theory will be shown in Chapter 13.)

7.6 Empirical Consistency: The Capitalist World

The model of general equilibrium with social disorder predicts that the sigma society operates with a higher degree of both income inequality and social compared to epsilon society. Therefore these should be the characteristics of the Third World countries with colonial legacy. Third World countries with weak colonial legacy would lie in between. The first part of the prediction is consistent with Fact 7 in Chapter 2.

The other part of the prediction is that social disorder will be higher in the Third World than in the First world. This is also consistent with the available empirical facts, as will be shown now.

According to the theoretical model presented here, there are several ways to measure social disorder. The most important measures found in the literature include property rights violations, political instability, and fragility of institutions.

Regarding property rights violations, an empirical study by the World Bank examined the statistical links between income inequality (measured by the Gini index from the data set constructed by Deininger and Squire (1996)) and violence (measured by crime rates from the UN *World Crime Surveys* data set). The sample includes 45 countries (16 from the First World and 29 from the Third World) for the homicide regression analysis, and 34 countries (14 from the First World and 20 from the Third World) for the robbery regression analysis, for the period 1965-1995. Both samples exclude countries from Sub-Saharan Africa due to lack of data sets. The main conclusion of this study is that income inequality has a robust and significant positive statistical association with the incidence of both types of violence (Fajnzylber, Lederman, & Loayza 2002). Bourguignon (2000) found the same statistical association on a sample of 50 countries, using data of crime rates for the period 1985-1995 and Gini index for 1985.

As to the prediction that higher income inequality leads to higher degree of political instability, Alesina and Perotti (1996), in a sample of 70 countries from the First World and the Third World, covering the period 1960-1985, found a positive correlation between income inequality and socio-political instability. The 16 Latin American countries in the sample had the highest degrees of income inequality and also the highest degrees of socio-political instability.

As to the prediction that higher income inequality leads to weaker democracies, political scientist Edward Muller, one of the most important scholars in his field, found a robust negative statistical relationship between the degree of inequality and the degree of democracy in a sample of 55 capitalist countries, which included 16 of the First World and 39 of the Third World, for the years 1960 and 1980 (Muller, 1997).

Regarding the prediction of weak institutions in very unequal societies, such as weak enforcement of the rule of law, the observed large size of the illegal activities, the so-called “informal sector”, in the Third World compared to the First World is also consistent with the theoretical model of social disorder. Illegal economic activities of large size are found mostly in drugs, mining, forestry, commerce (contraband) of people and goods. Government corruption should also be included here. In other cases, illegal industries are of small scale and supply “inferior goods”, that is, cheap goods for which the demand comes from the masses of low income people. The supply of inferior goods originates from robbery or piracy, and contraband; or from production at low cost, by escaping regulations and taxes. The illegal economic sectors play a significant role in general equilibrium: they generate employment (in the form of wage-employment and self-employment) and produce cheap goods and services that reduce the cost of living of the poor. Most illegal economic activities make a very unequal society viable.

First World countries have developed important systems of social protection. Individuals cannot go hungry because this risk is socially insured, as part of their economic rights. Take the example of unemployment insurance. Because the economy operates with

unemployment, the state has established a right to unemployment insurance. Thus unemployment does not generally lead to social disorder. The income compensation destined to the worker who is unemployed should receive a smaller amount than the market wage rate. If it were a full compensation, the labor market would be transformed into a Walrasian market, but the labor market cannot operate as such. In fact, this is how the unemployment insurance operates in the First World. Therefore, this way of compensation (a partial one) is consistent with the epsilon theory in which the role of unemployment is to assure labor discipline.

In the Third World countries, by contrast, systems of social protection are very weak and economic rights barely exist. Social disorder is usually controlled through authoritarian methods. Governments use repression measures to keep under control the political and economic violence (both collective and individual). Democratic rules cannot prevail when excessive inequality induces people to break some rules of the institutional context and authoritarian governments take over. The political system shifts from democratic to authoritarian governments continuously. This is, indeed, how democracy in a very unequal society functions in the real world, as predicted by the sigma theory.

The standard literature also recognizes the existence of weak institutions in the Third World. It should be noted, however, that the standard literature implicitly assumes that weak institutions are exogenous, as the major policy proposal is, for example, to have a strong judiciary system. By contrast, according to sigma theory, weak institutions are endogenous. Property rights violations (and the illegal sector), political instability, and degree of democracy, and a weak judiciary system, are all endogenous.

In sum, the findings of these empirical studies are consistent with the empirical predictions of the epsilon, omega, and sigma models when general equilibrium is with social disorder. These studies are consistent with the relationship shown by the curve R in Figure 7.3.

7.7 Conclusions

The theory of limited tolerance to inequality and the theory of government behavior have allowed us to construct general equilibrium models of epsilon, omega, and sigma societies, in which general equilibrium operates with social disorder. Societies can therefore, function with different degrees of income inequality and social disorder, depending on their initial inequality in the distribution of economic and political assets.

According with these models, we should observe that Third World countries operate with higher degree of both inequality and social disorder than is the case in the First World. This is indeed what the empirical studies show. The empirical data do not refute the predictions of the general equilibrium models of epsilon, omega, and sigma theories; hence, these models explain the functioning of capitalist societies. Hence, there is no reason to reject these theories at this stage of our research.

Unequal societies pay the price of inefficiency in the production process, as the potential output per worker and profits are not realized; another price is low quality of life due to being high risk societies. Therefore, inequality is not only a normative or ethical problem; it also has a role in shaping the efficiency and the quality of a society. In most of

the standard literature, inequality is seen as a normative problem only: too much inequality is morally unacceptable. However, the theory of limited tolerance to inequality says that it is a positive problem: too much inequality implies social disorder and thus lower quality of life for all, the rich and the poor. This theory also predicts that the capitalist system will lose legitimacy in a degree that depends upon the increase in the degree of inequality in society.

Regarding equilibrium prices in Walrasian markets, capitalism is a self-regulated system. Any excess demand or excess supply will be eliminated by the mechanism of market competition, as in the potato market. However, regarding excessive income inequality, which goes beyond the socially tolerable level, capitalism is not a self-regulated system. There is neither market mechanism nor democratic mechanism that can restore inequality to the socially tolerable region. The initial inequality in asset distribution (history) is the ultimate factor explaining the observed differences in income inequality and social disorder in different types of capitalist societies.

In sum, this chapter has shown that general equilibrium in capitalist societies is, by necessity, with excess labor supply. If general equilibrium tends to be with too high degree of income inequality, then society reaches the final general equilibrium with social disorder, which includes, by necessity, illegal economic activities. So far, these are the basic traits of the functioning of capitalist societies.

What role does the difference in the initial inequality between countries play in the process of economic growth? In particular, what is its role in the process of physical capital accumulation? This is the topic of the following chapter.

Figure 7.1. Region of Social Tolerance to Inequality

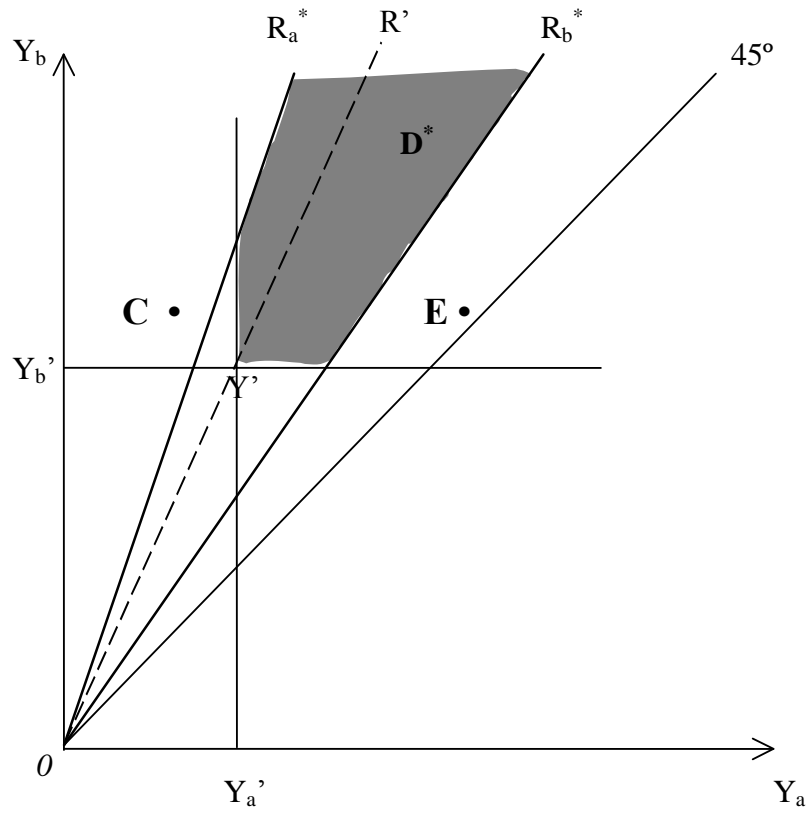


Figure 7.2. General Equilibrium with Social Disorder in the Capitalist Sector

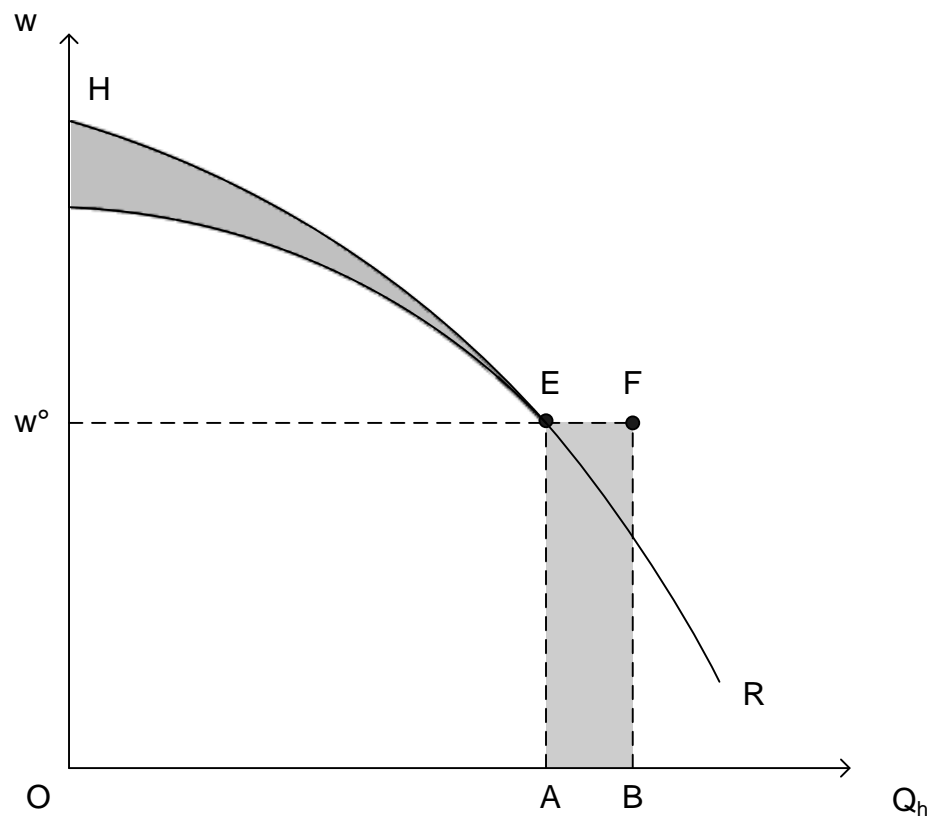
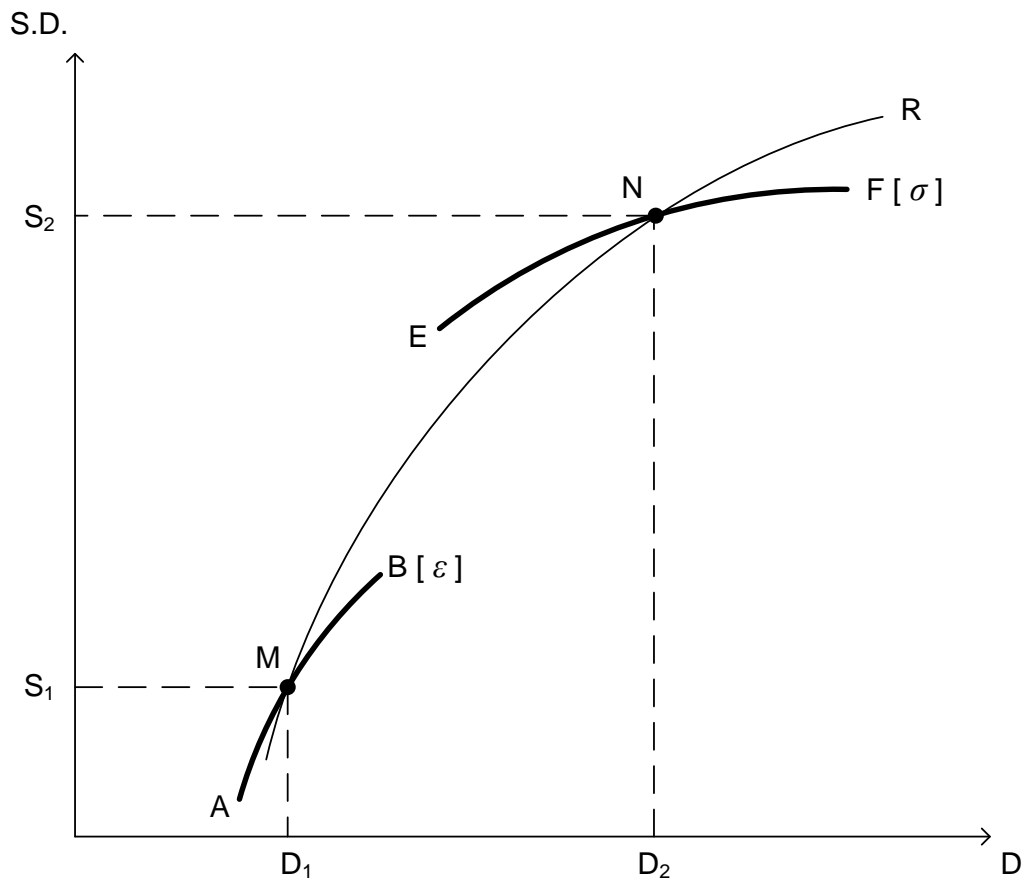


Figure 7.3. Relationships between Inequality (D) and Social Disorder (SD) by Types of Capitalist Societies



CHAPTER 8

INVESTMENT IN PHYSICAL CAPITAL

Standard economics has sought to explain investment differences in physical capital between countries by the differences in their factor endowments. The rate of return of capital is the fundamental determinant of investment and this rate is higher where capital is scarcer due to the diminishing returns of capital. Because the First World is more endowed with physical capital than the Third World, the prediction of this theory is that investment should flow from the former to the latter.

The prediction is that if investors wished to invest a given amount in the First World, they would get, say, 20% of profits; if the investment were placed in the Third World, the rate would be higher, say 40%. So investor would prefer to invest in the Third World. Empirical facts however, contradict this prediction. The bulk of foreign direct investment flows go mostly to the First World countries; only one-fifth of total foreign investment has been directed to the Third World in the last decades (Markusen, 2002; UNCTAD, 2006).

Professor Robert Lucas from the University of Chicago tried to save the theory by introducing human capital endowments into neoclassical models. First World countries are endowed not only with more physical capital but also with more human capital; therefore, diminishing returns of physical capital do not apply (Lucas 1990). In fact, the mean years of schooling of the labor force is higher in the First World compared to the Third World (Jones, 1998, Table B.2, pp. 180-183). This gap should be much higher if account is made of the higher quality of education in the First World. Thus the observed flows of capital could not refute the neoclassical theory.

The question still remains as to what the determinants of investment are. As a popular textbook on growth economics recognizes, neoclassical economics has not produced a “canonical theory” of investment in physical capital (Jones, 1998, p. 127).

Investors must choose not only projects, but also places where to invest. The investment theory to be developed here seeks to explain the allocation of investment to different regions of the capitalist world, which is composed of epsilon, omega, and sigma societies. The theory does not intend to explain the absolute level of investment, which is treated as exogenous. The predictions of the theory—which will be called the regional investment theory—will be confronted against the fact stated above, the low proportion of investment going to the Third World in the portfolio composition of investors.

Two other related questions are also analyzed in this chapter. The first refers to the implications of the investment theory for the competition between societies in the international commodity markets. The second is the role of credit and insurance markets in the process of accumulation of physical capital.

8.1 Aversion to Risky Games Behavior vs. Risk Aversion Behavior

The investment theory proposed here assumes that financial capital is perfectly mobile across societies. It also assumes that the investor's choice is taken in a context of uncertainty.

Regarding rationality, the primary assumption is the one stated before: capitalists seek two objectives: maintaining the social position and profit maximization, which are hierarchically ordered, the first has priority over the second. As investors, the rationality of capitalists will have to be consistent with the rationality just stated.

The theory about portfolio choice will be called the theory of aversion to risky games, which assumes the following rationality:

a(C).(3)Theory of portfolio choice: The investor's rationality to choose a portfolio is based on lexicographic preferences about rates of return and risk. This implies a sequence of decisions. First, the individual seeks to avoid portfolios where the risk of losses is unbearable; second, among the affordable portfolios, the investor will choose a portfolio according to this preference on rates of return and risk.

The first choice has to do with a class survival strategy; the second refers to the maximization motivation; moreover, the first objective has priority over the second. The exogenous variables include the resource endowment of the capitalist and the mean returns and risks of individual assets.

According to this theory, capitalists will seek to avoid playing economic games in which the risk of making losses under bad events could reduce the capital stock to a level below the threshold value to continue as member of the capitalist class, even if large gains might result under good events. If this loss happened, it would mean an economic disaster because the individual would stop being a capitalist. Therefore, capitalists set limits on the risk of losses to be taken; capitalists will play only those games where outcomes exclude any possibility of economic disaster. The existence of individual bankruptcies does not refute the theory because they are the exceptions and not the rule.

Under this rationality, firms are the means to obtain these objectives. According to this assumption, capitalists set up firms in particular regions to get profits, but have no interest in the firm itself or in the region per se. Thus, firms may go bankrupt, but capitalists do not necessarily disappear, for investment was chosen with a bearable risk of losses.

A model of the investment theory will be developed now. Some consistent auxiliary assumptions will be added. Firstly, in order to assure class position, capitalists will seek to have a level of wealth above a threshold. Let this threshold value be K^* . This is the minimum amount of capital that the individual must own in order to remain as a member of the capitalist class. Secondly, the portfolio choice refers to projects in capitalist societies, in which the rate of return and the risk depend upon the particular characteristics of each society. It is as if projects refer to invest in the production of the same good, but in different countries. The societies are two: epsilon and sigma. Thirdly, risk is measured by the standard deviation of the mean rate of return.

The typical investor's behavior can be represented by the following structural equations:

$$\begin{array}{ll}
\text{Given} & U = U [U_1 (1/S), U_2 (m, S)] \\
\text{Maximize} & U_2 = U_2 (m, S), S \leq S^*, U_{21} > 0, U_{22} < 0 \\
\text{Subject to} & m = f(S), f' > 0 \\
& m_j = g(K_{bj}, K_{hj}), g_1 < 0, g_2 > 0, \text{ and } j = \epsilon, \sigma \\
& S_j = h(\delta_j), h' > 0, \text{ and } j = \epsilon, \sigma \\
& R = \sum_j X_j \\
& S^* = F(R)
\end{array} \tag{8.1}$$

The first equation shows the lexicographic utility function of the investor, where m is the mean return and S is the standard deviation of the portfolio of projects. Let m_1 and S_1 represent the mean return and standard deviation of an investment in epsilon, and m_2 and S_2 the values corresponding to sigma; therefore, m and S will show the mean and variance of the portfolio of projects, which is equal to the corresponding weighted averages of those values, where the weights are determined by the mix of projects.

The second equation indicates that the capitalist seeks to maximize the second order utility function, subject to the constraint that the first order utility function is satisfied; here S^* is the threshold of tolerable risk, a deviation from the mean in units of the standard deviation. Thus, the threshold of income losses that are bearable by the individual is given by $(m - S^*)$ multiplied by the investment; beyond this threshold, the losses would imply economic disaster.

The third equation shows the feasible set of portfolios: the higher the standard deviation of a portfolio, the higher its mean return should be, if the two economies can compete in the portfolio of the investor. If investment in sigma has higher mean return than in epsilon, the risk must be higher in sigma than in epsilon; if risk were lower in sigma, then projects in epsilon could not compete and all the investment would go to sigma. More generally, portfolio choice requires that the society that presents lower rate of return should be less risky.

The fourth and fifth equations show the assumptions about the determinants of the rate of return and risk in each society. The rate of return in society j will depend negatively on the stock of physical capital (due to diminishing returns to the factor) and positively on the stock of human capital (infrastructure will have the same effect but for simplicity is ignored here). The risk of investing in society j will depend negatively on the initial inequality (δ). The assumption is that higher initial inequality implies higher income inequality, which in turn implies higher social disorder and much riskier society. It should be noted that the values of the rate of return and standard deviation are unobservable, for they depend upon expected subjective probabilities attributed by investors to the projects in each society. But these relations indicate what observable variables depend upon.

The two final equations in the system (8.1) indicate idiosyncratic restrictions that the investor faces. The first says that the investor is endowed with a given amount of investment fund R , which must be allocated to these two economies. The second says that the investor's bearable losses depend positively upon his wealth. Then big investors will be able to invest in very risky project, in which small investors will not.

If the possible losses would leave the capitalist with wealth equal or higher than the value of capital needed to be a member of the capitalist class (K^*), then we say that those

losses are bearable, do not generate economic disaster. Suppose a portfolio “A” with mean return of 10% and risk of losses of 20% and a portfolio “B” with mean return of 30% and risk of losses of 50%. Let the value of K^* be equal to 100 dollars (in thousands or millions to make it realistic). If the capitalist’s wealth is 150 dollars, choosing portfolio “A” implies a possible loss of 30 dollars (20% of 150) and a possible reduction of wealth to 120, which is still higher than 100 dollars. Choosing portfolio “B” implies a possible loss of 75 dollars (50% of 150) and a possible reduction of wealth to 75, which is below the threshold of 100 dollars. Hence, the capitalist will choose portfolio “A” instead of portfolio “B”, even though it has a higher mean return (and a larger gain if the good event occurs). Suppose the capitalist has a new wealth of 300 dollars, then he could consider choosing portfolio “B” because the possible loss is 150 dollars, which is bearable because the new wealth under this unfavorable outcome would be 150 dollars.

From this example, it is clear that the higher the capitalist’s wealth is, the higher the total losses he can bear, and the higher the mean return he can get from investments. In other words, in order to choose very risky portfolios, the capitalist must be very wealthy. The implication is that capital accumulation tends to increase the wealth inequality within the capitalist class because the very wealthy, relative to the less wealthy, will be able to invest in very risky portfolios but with the possibility of obtaining high mean returns, which will increase further his wealth.

Capitalists also face the risk of destruction of their capital stock in every period. This risk can be insured through the insurance market. The capitalist will purchase an amount of dollars of insurance and will pay a premium per unit of time. The physical capital insured will not be smaller than K^* . In other words, the capitalist will not seek to play risky games, that is, games that might lead to a loss beyond the bearable level, even if large profits might occur if the events were favorable. The capitalist behavior is guided by the motivation of *aversion to risky games* in this precise sense.

This assumption is different from *risk aversion*, which is the assumption made in standard economics. The capitalist will prefer a portfolio that offers a higher mean return at the same risk, or lower risk at the same mean return; but the capitalist will also prefer a higher risk portfolio if the mean return is sufficiently high to compensate for the high risk. There are no limits to risk under the assumption of risk aversion behavior. The capitalist does not care for economic disaster.

Risk aversion behavior predicts that the portfolio will not change as wealth increases. An optimum portfolio remains optimum if the investor has more capital to invest. The aversion to risky games behavior predicts that the portfolio will change as wealth increases. As the wealth of the investor increases, he can take more risky portfolios because this higher risk was not bearable before, but it is now. In the real world, portfolios seem to vary across wealth levels of capitalists. This fact clearly refutes the prediction of the risk aversion theory, but it does not the prediction of the aversion to risky games theory.

In sum, the investor’s portfolio choice will depend on the differences in the rate of return and the risk of projects between epsilon and sigma societies, which in turn depend upon differences in their factor endowments and initial inequality. On the rate of return, the difference is ambiguous because, by assumption, sigma is endowed with lower stocks of both physical capital and human capital compared to epsilon; the first is favorable but the second is unfavorable. On the standard deviation, the difference is clear: risk is higher in

sigma than in epsilon because, by assumption, the degree of initial inequality is higher in sigma than in epsilon.

The equilibrium condition in the individual investor's behavior can then be stated as follows: the optimum portfolio depends on the differences in the mean return only. If the mean return in sigma is lower or equal than in epsilon, all investment fund will go to epsilon; if the mean return is higher in sigma, then equilibrium will imply a portfolio choice, part of the investment fund will be allocated to epsilon and the other part to sigma.

8.2 Aggregate Investment Behavior

Aggregating the individual behavior over all investors, the *proportion* of the global investment in physical capital (good B) going to sigma society can be represented by the following function:

$$i_b(\sigma) = f(K_{b\sigma}/K_{b\epsilon}, K_{h\sigma}/K_{h\epsilon}, \delta_\sigma/\delta_\epsilon), f_1 < 0, f_2 > 0, f_3 < 0 \quad (8.2)$$

The term $i_b(\sigma)$ indicates the proportion of total world investment flowing to sigma economy; K stands for the stock of capital and δ is the initial inequality; sub-indices b and h stand for physical capital and human capital, whereas σ and ϵ indicate sigma and epsilon economies. The other proportion of global investment will go to epsilon.

Beta propositions of this model are given by the partial derivatives of the equation. Investment flows going to sigma society will be higher, the lower is the relative endowment of physical capital, the higher is the relative endowment of human capital, and the lower is the relative initial inequality compared to epsilon society. The proportion $(1-i_b(\sigma))$ will go to epsilon society.

The *level* of investment going to sigma society will be just the proportion of investment multiplied by the total investment. Then this level will depend upon the same exogenous variables indicated for the proportion of investment, to which the global investment fund should be added.

The global investment fund is exogenous. It includes all the financial sources that investors can use to carry out their investments, which include business profits, loanable funds in the credit markets, and funds in the financial markets. As this fund increases, the individual investor will increase his level of investment, but the portfolio will not change. The reason is that the mean and standard deviation of projects depend on factor endowments and the initial inequality alone.⁷

Profits remittances across countries require foreign exchange. Until now, the epsilon, omega, and sigma models assumed that firms were all national; then no profits had to be remitted abroad. But even if some firms were foreign and profits had to be remitted abroad, the one-single economy model solved this problem easily: part of the real output

⁷ In the standard economics literature, we find an investment theory in which private investment depends on the interest rate and relative prices of competing assets (cf. Barro, 1997). It should be noted, however, that this is a theory of investment level, about the variable X; it is not a theory of the allocation of investment among different types of capitalist societies.

went abroad as profits. In a two-good model, which is necessary in an international model, the firm that is producing one commodity will need to purchase foreign exchange to remit profits abroad. Balance of payments problems may then result from profit remittances. This problem will be ignored in this model. However, profit remittances affect income inequality between countries. If most investment originates in epsilon societies, then profit remittances from sigma to epsilon will increase income in epsilon and thus, income inequality between epsilon and sigma will go up.

A prediction of the investment model refers to the observable equilibrium conditions. It says that the proportion of the total investment flow going to any one capitalist society (epsilon, omega, or sigma) will be higher, the lower is its relative endowment of physical capital, the higher is its relative endowment of human capital (and infrastructure capital), and the lower is its relative initial inequality compared to the other two societies. Consider the proportion going to sigma society. Given that sigma society is relatively less endowed of both physical capital and human capital than epsilon, the effects of factor endowments tend to cancel out (the effect of the first is positive and that of the second is negative). Therefore, the only important factor comes from differences in “inequality endowments,” which is negative and predicts that the proportion will be small.

Another prediction refers to changes in the exogenous variables. Changes in the proportion of investment going to any one of the capitalist societies will depend upon changes in the same exogenous variables. The effects are indicated by the signs of the partial derivatives shown in equation (8.2).

A third prediction is that investment going to sigma type societies will be made by large investors. Risk is higher in sigma societies and only very wealthy investors can bear the risk of large losses. In the aggregation, investors having different wealth sizes, and able to absorb different levels of risk, are being combined. Big firms will be able to invest in high risk projects, in which small firms will not. Therefore, investors who are going to invest in sigma societies (high risk society) and in projects that are also very risky (oil, gas, minerals) will be mostly big investors.

8.3 Empirical Consistency: The Capitalist World

On private investment in physical capital, the theoretical model presented above has produced a set of empirical predictions. Given the availability of empirical studies in the international literature, the consistency of the predictions with facts can be presented as follows.

The study of Alesina and Perotti (1996) referred to above found not only a positive correlation between income inequality and political instability (as reported in Chapter 7), but also a negative correlation between political instability and private investment. In their statistical analysis, they use the distribution of income of the initial period of the sample, 1960, which may be a good approximation of the initial inequality of assets; for the other variables, they use the mean values over the period 1960-1985. The interpretation made from their results is that “income inequality increases socio-political instability, which in turn decreases investment” (p. 18). Moreover, the variable “income inequality” actually means “initial inequality” because the empirical variable utilized refers to income inequality in 1960. This result is consistent with the prediction of the investment model.

A study on the behavior of multinational corporations proposed a theoretical model on investment determinants, which predicts that investment depends positively upon infrastructure capital and human capital of countries (Markusen, 2002, p. 212). Income inequality is not included in this model. Empirically, the study indeed finds a strong and positive statistical relation between human capital and investment level made by those multinational enterprises operating in First World and Third World countries at the same time. This result is also consistent with the prediction of the investment model.

On the prediction about the proportion of investment, empirical studies show that the bulk of foreign direct investment (75%) flew to the First World countries in the decade of the 1990s (Markusen 2002, Table 1.2, p. 9). The UNCTAD study also shows that of the foreign direct investment stock in 2004, the share of the Third World was only 22% (UNCTAD 2006, Table 3.9, p. 108). From the figures presented in the annual World Bank Development Report, the following proportions can be derived: 75% was allocated to the First World and 25% to the Third World in 2009, whereas these proportions were 80% and 20% in 1990 (World Bank 2001, 2010, Table 21).

Although some shift may be taking place over time, as the World Bank data suggest, the level of the proportions is clear in showing that most of direct foreign investment originates in the First World and also goes to the First World. Markusen study indeed concludes by saying, "Not surprisingly, the developed countries are the major source of outward investment, but perhaps less well known, they are the major recipients as well" (p. 8). On the distribution of profits between the groups of countries, the implication of the investment allocation is also clear: profits originate mostly in the First World and are received mostly in the same First World.

The fact that the proportion of direct foreign investment going to the Third World (sigma societies) is smaller than that going to the First World (epsilon societies) is consistent with what the investment model predicts. Differences in factor endowments and initial inequalities between the First World and the Third World indicate that the former is more endowed with physical capital and human capital and that its initial inequality is relatively lower. Under these conditions, the effects of factor endowments on mean returns tend to cancel out; then risk is the relevant element in investment decisions. Risk is higher in the Third World and few investment projects will have a mean return that is higher enough in the Third World to compensate for the high risk and be able to compete in the investors' portfolios.

The last prediction is that direct foreign investment going to the Third World (sigma societies) will be carried out mostly by large firms, those that can bear large losses due to the high risk. Empirical studies are needed to see whether large multinational corporations, investing in the exploitation of natural resources, is the main feature of foreign direct investment in the Third World.

In sum, the model of the theory of investment presented here predicts that factor endowments and the initial inequality play a significant role in the process of physical capital accumulation of capitalist societies. Available empirical facts do not refute these predictions; hence, there is no reason to reject the investment theory and we may accept it at this stage of our research.

8.4 Inequality and International Competition

The standard neoclassical model of international trade theory predicts that labor abundant countries will specialize in labor-intensive industries and that trade will lead to real wage equalization across countries. Trade of commodities is thus a perfect substitute of factor mobility. Exporting goods that are produced with a significant amount of labor content is equivalent to the emigration of workers: both imply labor scarcity and higher real wages. Facts have contradicted these predictions; in particular there is no real wage equalization among trade partners. Trade of commodities is not a perfect substitute of factor mobility as the theory assumes.

The neoclassical theory of international trade ignores the role of inequality in the economic process. Inequality plays the role of a negative factor of production due to the social disorder it generates, as shown in Chapter 7. A new concept of factor intensity thus appeared: some industries will be relatively equality-intensive compared to others, which is similar to the concepts of relatively labor-intensive industries or relatively capital-intensive industries. The productive process of a good is less equality-intensive when it may be produced in isolation for it requires little sectorial inter-linkages. This will be the case of investments aimed at producing goods in enclaves, such as mines, oil fields, sweat shops, and tourist centers. The society might be in a big social turmoil, but the production and investment activity in these sectors and places will continue mostly undisturbed. Social disorder is not too costly to this industry.

A different pattern of specialization in international trade can then be predicted from the theory of investment presented here. Sigma society is a labor abundant and should specialize in labor-intensive industries and should attract investments in these industries. However, labor-intensive industries are also equality-intensive industries and thus labor-intensive industries will lose the expected comparative advantage; that is, capital-intensive industries, which are not equality-intensive, which can be produced in enclaves, will gain relative competitiveness. Only a small fraction of private investment will then be directed to the exploitation of labor-intensive industries.

Therefore, another empirical implication of the model is that sigma economies will not specialize in labor intensive commodities in international trade. This prediction is consistent with the empirical fact that international trade has not significantly reduced overpopulation in the Third World. In these countries, foreign investment is mostly directed to exploit natural resources such as minerals, oil, gas, tropical products, or tourist sites, which are generally produced in enclaves.

There is an effect of the “inequality endowments” of the economy upon its industrial structure. Sigma economies will not be as industrialized as epsilon economies. This is consistent with facts. The First World is more industrialized than the Third World, whereas the latter is more specialized in natural resource exploitation.

These results indicate that countries compete in two arenas in the international economy: in the commodity markets and in the capital markets (investment in physical capital). But competition in the capital markets, attracting private investment, is prior to competition in the commodity markets. In both cases, countries compete not only with their factor endowments, but also with their “inequality endowments,” the other component of their initial conditions.

In this investment theory, the only investors are the capitalists. For one thing, they have their own profits to finance investments. Also they can borrow from financial markets, through loans from banks or through emissions of bonds or shares. Why is it that workers do not accumulate physical capital? The usual answer is lack of savings. But investment can be financed through loans from the credit market. Workers would then need to have access to the credit market in order to finance the accumulation of capital; they would also need access to the insurance market to protect the capital from destruction. The fact is that workers do not have access to these markets. The study of the nature of credit and insurance markets is essential to answer the question. Let us then, turn to this matter.

8.5 The Role of Credit Markets

Include now the credit market into the epsilon, omega, and sigma models. Individuals may now borrow money from banks in order to finance capital accumulation, either to expand existing firms or to set up new firms. Because in a credit market people exchange a sum of money against the promise to repay it, standard economics assumes that the credit market operates under a context of asymmetric information. Individuals are better informed than banks about their credit worthiness. These two parties' incentives are incompatible, and default is possible. Banks and borrowers have a principal-agent type of problem, similar to the firm-workers relations studied above. These assumptions are common in standard economics (cf. Stiglitz and Weiss, 1981).

Additional assumptions will be introduced here to construct a very simple credit market model that is able to explain the observed exclusion from access to credit markets. Suppose the credit market is Walrasian and operates under perfect competition. Define demand for credit as the quantity of money that individuals are willing to borrow and able to repay at the market interest rate. Suppose the demand curve is downward sloping: the lower the interest rate, the greater the quantity of credit demanded. This is the result of potential investment projects (in need of financing) that yield different rates of return; therefore, the number of potential investment projects that yield at least 5% will always be greater than the number of projects that yields at least 15%.

Suppose the supply curve of loanable funds by banks is upward sloping; the higher the interest rate, the greater the quantity of loanable funds supplied to the market. Banks are intermediary firms between people with excess of funds and those with deficit of funds. The market upward sloping curve comes, in part, from the upward supply curve of bank deposits, which originates in those units with excess of funds, say the workers. Workers' demand for money now includes cash balances and bank deposits that yield interest. The opportunity cost of cash is the bank's deposit earnings.

The bank can reduce the asymmetric information problem by incurring in transaction costs; that is, by obtaining information about the debtor and the project. Transaction cost is also assumed to be upward sloping, and constitutes the other component of the variable cost of banks. Given that the credit market is Walrasian, there will be an interest rate that clears the market.

But, can the credit market operate this way? What would be the economic incentives for borrowers to repay the loan? If an individual can always get credit, independently of his history of repayments, why should he bother to repay the loan? Borrowers will not suffer

any cost if they default. The rate of repayment would be very low and banks would go bankrupt. Therefore, the credit market could not operate as a Walrasian market.

Banks have to use devices to reduce the risk of dealing with bad borrowers and bad projects in order to ensure repayment. A possible device would be to set the interest rate below the Walrasian price. Excess demand would thus be generated and the credit market would operate with quantitative rationing among borrowers. The cost of default would mean exclusion from the credit market. Opportunistic behavior of borrowers could not be repeated and default would thus be deterred with this device. In a static economy, however, the borrower may build up some capital by taking credit only once and not repaying the loan. In addition, the setting of this interest rate in the market needs collective action among banks. Therefore, this device would not work well.

Assume now that banks use collaterals as the particular device to reduce the risk of default. Suppose that the amount of the collateral is set equal to the amount of the loan and that in the case of defaults, the collateral can be executed at zero enforcement cost. The loan is thus fully insured. Heterogeneous borrowers have thus been transformed into homogeneous borrowers in terms of risk.

But now individuals who have no capital at all and are willing to borrow at the market interest rate—and having potential investment projects that yield returns at least as high as the interest rate, and thus able to repay the loan—will be excluded from the credit market. This device therefore causes a downward shift of the credit demand curve.

Would all individuals endowed with capital, no matter how small, be able to obtain bank loans? Suppose that the transaction cost *per borrower* is constant; that is, it is independent of the loan size. The cost of evaluation, contract, disbursement, monitoring, supervision, collecting the repayment, and enforcing the collateral does not depend on the loan size; that is, the transaction cost per borrower is a fixed cost. What we are saying is that intermediation in the financial market will be economically profitable in the presence of fixed transaction costs. This assumption implies that the unit cost *per dollar* of the loan declines with the loan size. If the total cost of supplying a loan of 100 thousand dollars is the same as that of a loan of one thousand dollars, the bank will seek to give one loan of the first size rather than 100 loans of the second. There are economies of scale in the supply of loans. The bank will seek to maximize profits, which implies that it will avoid small size loans.

Assume now that the unit cost per dollar does decline up to a certain loan size and then remains constant; that is, there are limits to the economies of scale. This assumption allows us to establish a threshold value for the minimum size of bank loans.

The bank's set of feasible profits will then be higher with large size loans than with small size loans. Banks will not engage in supplying small size loans, smaller than this threshold loan size; thus, retail bank loans will not exist. Consequently, banks will supply loans of sizes equal or larger than the threshold value, which will determine the corresponding threshold value for the capital endowment of borrowers. Let the amount of capital K^* represent this threshold; that is, to be eligible to apply for a credit, the individual must be endowed at least with K^* units of physical capital.

In the aggregate, the market demand for credit will not be determined by all individuals endowed with physical capital and willing to get credit, but only by those

individuals who own capital in an amount that is above the threshold value K^* . To distinguish from total demand for credit, call *effective demand for credit* the demand for credit coming from those borrowers that are eligible by the banks. It is clear that market demand and supply are not independent as is the case of potato markets. The effective demand for credit does not show the behavior of borrowers alone, but also the behavior of lenders. Hence, if the collateral threshold decreases, then the effective demand level will increase, that is, there will be more borrowers in the market, even though the total number of potential borrowers remains unchanged. The exclusion mechanism based on wealth is thus embedded in the effective demand for credit. This model predicts that the poor are excluded from the credit market.

In this model of credit market, the interest rate will clear the restricted market, equalizing the quantities of the effective demand with that of the supplied. This is illustrated in Figure 8.1, at point G. Is the credit market Walrasian? No, because at the equilibrium interest rate there will be excess demand. Total demand is given by curve D_0 , which includes people with capital endowments above the threshold value (curve D_2), below the threshold value (difference between curve D_1 and D_2), and those with no endowments (difference between D_1 and D_0).

At the equilibrium interest rate r^0 there are individuals who are willing to borrow and able to repay but are not eligible for credit (very similar to the way the labor market operates). Not everyone willing to exchange the quantities of credit they desire at the prevailing market interest rate will be able to do so; therefore, the credit market would be non-Walrasian. But at the equilibrium interest rate no social agent has the power and the will to change the situation; thus, this is an equilibrium situation. The effective demand constitutes the relevant notion of demand, and with respect to this the credit market operates as a Walrasian market. The credit market operates *as if* it were Walrasian. Hence, it may be better to define the credit market as *quasi-Walrasian*.

In standard economics, no canonical model of the bank credit theory exists. Macroeconomic models assume that the credit market is Walrasian (just as the potato market) (Barro 1997, chapters 5 and 17; Krugman and Wells 2006, chapter 9). Some models include the assumption of risk aversion in the behavior of borrowers; thus at given collateral requirements, some individuals may decide not to borrow even if they have the necessary capital due to the risk of losing their assets (Blinder, 1987). But in this case people self-exclude voluntarily from the credit market. Then the credit market is Walrasian.

Among microeconomic models, some show that credit markets *could be* Walrasian, some that it could be non-Walrasian (equilibrium with credit rationing), and others that market equilibrium may not exist (Freixas and Rochet 2008, chapters 4 and 5). These microeconomic models are then unfalsifiable, as their predictions are not definitive observable situations but just possibilities. Popperian epistemology says that “could be” predictions of a model make the model immortal.

Two empirical regularities that are observed on the behavior of bank credit markets in the First World and Third World are the following: collateral use is the common practice and there are no retail bank loans. The latter implies exclusion of the potential borrowers that are poorly endowed with physical capital. These two facts are indeed consistent with the predictions derived from the particular model of the standard theory of the credit market presented here. Those facts are the equilibrium conditions under which banks operate.

A third characteristic of the bank credit market is its coexistence with other forms of credit exchange. This follows from the bank credit model. If bank credit market operates with excess demand, where does the excess demand go for credit? In the Third World the other forms of credit comprises the non-banking formal industry and the informal sector. This coexistence needs a theoretical explanation.

Inequality in the endowment of physical capital is more pronounced in the Third World than in the First World. Therefore, the excess demand in the bank credit market will be more significant in the Third World than in the First World. Those excluded from the bank credit market seek small size loans in the formal non-banking credit organizations (cooperatives, loan and deposit associations). In this formal market, it is known that the loan interest rate is higher and the market size is much smaller than in the bank credit market. The reasons for the higher loan interest rate are the small size of loans and the higher deposit interest rate that must be paid to attract deposits that otherwise would go to the banks.

The non-banking sector does not operate as a Walrasian market either. Those excluded from this market will constitute another segment of the excess demand, who will seek loans in the informal sector, in which the loan providers include individuals (including here the “loan sharks” as well as friends and relatives) and some non-profits organizations, such as NGOs. One of the characteristics of the informal credit sector is that it is not state regulated. Credit contract are thus informal. The other is that interest rates are higher than in the non-banking credit market and the borrowers are the poorest in society.

Theoretically and empirically, it has been shown that there is a connection between these forms of credit exchange. The equilibrium prices and quantities in the banking industry determine the prices and quantities in the formal non-banking industry, which in turn determine the prices and quantities in the informal sector. This dual-dual financial structure that we observe in the Third World, but not in the First World, is explained by its relatively more pronounced inequality in wealth (Figueroa 2012).

In sum, in sigma societies, the poor are excluded from the banking industry and are thus limited to accumulate physical capital. Formal and informal sectors cannot offset the cost of this exclusion. In the non-banking formal sector the excluded will be able to obtain loans of small size, but they will be more expensive and limited to few people due to the limited total quantity supplied. In the informal sector, the poorest of the excluded will seek loans, which are the most expensive, the smallest in size, and for the shortest periods. In sum, the poor faces the greatest disadvantage to accumulate physical capital.

The empirical predictions of the credit model presented here are consistent with the three basic facts of the real world. Therefore, there is no reason to reject the credit model at this stage of our research.

8.6 The Role of Insurance Markets

Uncertainty is a prominent feature of the world that we humans live in. In such a world, individuals make decisions under different contexts of uncertainty, called states of nature or contingency situations, which depend upon whether risk (the likelihood of suffering an economic loss) is measurable or not and also whether the risk is bearable or not by

individuals. The insurance market can operate only in the context in which individuals face a risk that is measurable and unbearable.

The Nature of the Insurance Market

What is the nature of the insurance market? What do people exchange in this market? The good exchanged in the insurance market is a very special one: the insurance firm promises to pay out for an accidental damage on the object insured and the individual in return pays an annual sum of money, called the premium, which is equal to price (premium rate) multiplied by the quantity exchanged. Because people exchange promises, contracts are utilized as the means of exchange. This market exchange is intended to make the insured object indestructible, either totally or partially, depending on the contract policy.

Insurance firms will in turn be willing to supply insurance contracts to the market because in the aggregate risk will be measurable and, therefore, cost and profit calculations become feasible. The firm can make those calculations because the large number of buyers makes possible to calculate the probability of the aggregate damage by using the principle of the statistical law of large numbers. If a fair coin is tossed once, and “head” implies losses, then the loss has probability of $\frac{1}{2}$; if 4 coins are tossed, the probability of getting all “heads” is only $\frac{1}{16}$; if 10 coins are tossed this probability is $\frac{1}{1,024}$; imagine the probability of total losses if 100 coins are tossed!

If the firm insured just one individual, the firm would be taking the risk of the individual. But in a market exchange, with many buyers, the probability of losses can be calculated with a higher degree of accuracy (the statistical theorem shows that larger samples generate smaller variance). The average damage is then known. Therefore, the risk facing individuals are transferred to the insurance firm; but in the firm it tends to disappear with the aggregate of buyers. The insurance firm can then do business and seek profits with a degree of risk that is just similar to that in other industries.

The basic assumption of the standard theory of insurance markets is that insurance markets operate in a context of asymmetric information. The insurance market faces two problems, referred to as “adverse selection problem” and the “moral hazard problem.” The first refers to the problem that the risk attributes of individuals who seek to buy insurance are diverse and unknown to the insurance firm or the information is imperfect; the firm thus runs the risk of ending with a collection of the worst buyers, which would increase payments for accidental damages and would reduce profits.

What would insurance firms do in order to eliminate or reduce the adverse selection problem? Because individuals or firms seeking to buy insurance are heterogeneous in their degree of risk, insurance firms incur in transaction costs in order to know better the risk situation of prospect buyers.

The moral hazard problem refers to the problem of changes in the behavior of individuals once they have purchased the insurance, the effects of which on the size of the average accidental damage is to increase it. The assumption is that the purchase of insurance changes the behavior of the buyer, who now has the incentive to act carelessly in comparison to the situation of no-insurance. In the event of a shock on his endowments, the uninsured buyer bears the full cost of the damage; however, if it is insured, he will pay only

a fraction of that cost and, hence, the incentive is generated for acting carelessly when insurance is bought.

What would insurance firms do to eliminate or reduce the moral hazard problem? The insurance contract can hardly specify all the actions that the buyer is to undertake and the required supervision would be costly. But the theory assumes that the insurance firm will use an incentive device that induces the buyer to behave in ways that he or she will bear the consequences of their careless actions beyond the price of the insurance. These devices can take the form of co-payments clauses and deductibility provisions in insurance contracts. The cost of sharing damages could now discourage careless behavior of those who purchase insurance.

Insurance firms can charge different prices to different buyers because arbitrage is difficult in the insurance market. Moreover, the discrimination devices will tend to increase the total price of the insurance for high risk buyers and thus may also discourage high risk clients from buying the insurance in the first place. Hence, insurance firms use devices to generate incentives against both adverse selection and moral hazard problems; thus insurance firms may seek profit maximization in ways that are similar to what firms of other industries follow.

On the individual buyer, the standard theory assumes that the individual facing an unbearable risk will be willing to buy an insurance contract, at a price that is a small fraction of the value of the asset because then the risk will disappear for him. The individual seeks to make this transaction because he prefers a situation in which there is a secure small loss (the annual expenditure in buying insurance) and a small chance of a large gain (the insurance payout after disaster) to the alternative situation in which there is a secure small gain (no premiums) but a risk of a huge loss (no payout after disaster). Individuals will then be willing to buy insurance to avoid unbearable losses, losses that can imply an economic disaster. The individual can thus transfer the risk to the market.

Property Insurance Market

The market for insuring physical capital is the main objective in our analysis of investment. To this particular insurance market we now turn. A very simple model will be developed.

Insurance firms will incur in transaction costs as a device to reduce the problem of asymmetric information, as we pointed out above. Assume that the transaction cost per contract is fixed, independent of the quantity of insurance that the buyer will buy; therefore, the cost *per dollar* of insurance sold is a rectangular hyperbola curve until a threshold size is reached, beyond which it becomes constant. The transaction cost shows economies of scale up to a certain size of the insurance sold.

Let the threshold size be x^* dollars of insurance. Thus the insurance firm will be willing to sell only insurances sizes that are above this threshold. Insurance firms seek to maximize profits and, consequently, will have no incentives to sell insurances of smaller size. Insurance firms will avoid engaging in the sale of insurance at retail. The existence of economies of scale in transactions of large size insurance is in the nature of transactions in

the insurance firms. Technology may change the shape of the curve, but the economies of scale will remain.

The standard cost analysis of firms usually makes no distinction between the total quantity supplied and the sizes of the parcels that are exchanged of that total. The latter category is usually ignored. However, in the exchange of some goods (such as insurance and credit) this distinction is essential due to the existence of economies of scale in the transaction cost per size of the parcel. Therefore, firms must choose not only the optimal total output to supply to the market, but also the optimum sizes of parcels to be exchanged with individual buyers.

In what follows, therefore, the model assumes that insurance firms choose, firstly, the optimum sizes of insurance and secondly the optimal total quantity of insurance, the one that maximizes profits, subject to marginal costs, price of insurance prevailing in the market, and also subject to the constraint that sizes of insurance must be higher than the threshold value (x^*).

Insurance firms use the funds raised through premiums to invest in financial markets and earn financial income. They constitute the so-called institutional investors and have a significant role in financial markets. But in order to simplify further, the model will assume that financial decisions are sequential. First insurance firms choose price and quantities to be insured and then decide on financial income. So the model will intend to explain price and quantity determination in the insurance market and ignore the financial decisions of insurance firms.

Consider now the behavior of the individual buyer of insurance, which will be the individual capitalist who owns a firm producing goods. The capitalist is endowed with stocks of physical capital in the amount K_{b0j} , measured in units of good B, which is the only capital good produced in the economy. From this capital stock, the capitalist generates income in the form of profits (P). In the world of uncertainty in which the capitalist operates, assume just two possible events: bad and good, with known probabilities π_1 and π_2 . In the bad event case, assume the risk is that the capital can be totally lost; thus the capitalist's initial endowment is $(K_{b0j}, 0_j)$, and the corresponding income endowments are $(P_j, 0_j)$, in the good and bad states.

In the bad state, therefore, not only the individual capitalist losses the capital stock and thus profits will be zero, but the individual capitalist will have to abandon the privilege position of being a member of the capitalist class. If the bad event occurs the individual will thus suffer an economic disaster, which he will seek to avoid.

What choices does the individual capitalist have when the insurance market is available? Buy buying insurance in the market, and assuming his capital endowment is fully insured, the capitalist can transform the initial capital endowment into (K_{b0j}, K_{b0j}) , and the corresponding initial income endowment into $(P_j - C, P_j - C)$, where C is the premium total cost. The machines can thus become indestructible through market exchange. Because firms use depreciation cost to maintain the same stock of physical capital over time, depreciation makes capital durable. Now when insurance is bought, the stock of capital becomes durable and indestructible. Production and profits net of premium total cost can be repeated period after period.

The aggregate market behavior can now be established. Assume perfect competition in the insurance market for property insurance. Insurance firms will be willing to supply more quantity of insurance at higher price. Capitalists will be willing to buy more insurance at lower price. But that is not all. Not everyone willing to buy insurance can do it. Only those capitalists endowed with a stock of capital that is higher than the threshold valued established by insurance firms ($K_{bj} > K_b^* = x^*$) will be able to participate in market exchange. The device to minimize costs of transaction utilized by insurance firms implies the exclusion of capitalists with low capital endowments from the insurance market.

Is the insurance market Walrasian? Supply and demand are not independent, as insurance firms determine the eligibility of people who can participate in the market. The segment of the total demand that effectively participate in the market will be called the *effective demand for insurance*. An exogenous reduction of the threshold value will imply a higher level of demand, even though the total demand level remains constant.

Price and quantity of equilibrium in the insurance market will be determined where the effective demand curve and the supply curve cross each other. Figure 8.1 can now be seen as representing the insurance market, just by changing the definition of prices and quantities in the axes. At the point of intersection, point G, there will be excess demand for insurance. Then this equilibrium would imply a non-Walrasian market. The effective demand curve (not the total demand curve) is the relevant one in the functioning of the insurance market and thus, the interaction of the effective demand curve with the supply curve will determine the price and quantity of equilibrium. This is an equilibrium situation because no one has the power and the will to change this solution. In this case, the insurance market operates *as if* it were a Walrasian market. Therefore, the insurance market may be defined as *quasi-Walrasian*.

Once equilibrium has been reached in the insurance market, the price and the quantity of equilibrium will be repeated period after period, as long as the exogenous variables remain fixed. The static market equilibrium is with exclusion of the less wealthy capitalists, who are unable to buy insurance to protect their physical capital from the risk of destruction.

The existence of the insurance market allows those capitalists that have access to this market to transform a situation of unbearable risks into a situation of bearable risks, at a fixed cost, the cost of buying the insurance policy. Furthermore, the insurance market allows the capitalist to have an indestructible capital endowment, which ensures membership of the capitalist class, although at the cost of lower profits (P-C). But this is consistent with the primary assumption that capitalists have a hierarchically order preferences, in which remaining as part of the capitalist class has priority over the objective of profit maximization. Capitalists with small capital endowments are excluded and subject to economic disaster.

Two empirical regularities that we observe in the insurance markets of the real world include:

- No retail market is observed in insurance markets; individuals with small endowments of physical capital are excluded from the insurance market.
- Insurance markets usually operate with co-payments and deductibles in addition to the price of the insurance.

The theoretical model presented here predicts these features. The model cannot be rejected at this stage of our research.

8.7 Basic Markets

If the credit and insurance markets were integrated into a general equilibrium model of epsilon, omega or sigma theories, the market system would be constituted now by six markets, of which three are Walrasian (good B, money, and foreign exchange), one is non-Walrasian (labor), and two are quasi-Walrasian (credit and insurance).

No attempts will be made here to present the general equilibrium solution of these models. The solution of each general equilibrium model will involve more interactions. It suffices to say that the credit market introduces into the system two additional endogenous variables (price and quantity) and two additional conditions or equations (supply and effective demand). Therefore, the theory of markets is still valid: the market system operates as if it solved a system of equations.

The introduction of the insurance market is much simpler. It is solved sequentially, under a partial equilibrium analysis, after the general equilibrium solution has been determined. The reason is that the quantity of insurance bought by capitalist firms is a fixed cost and therefore, it will not affect the behavior of the production firms, except that profits net of insurance cost will be lower.

It also suffices to say here that the new equilibrium will be stable and that comparative statics can be applied to derive beta propositions. And that, because the exogenous variables of the new general equilibrium model are the same as in the old one, the new model predictions are still consistent with the relevant empirical regularities, Facts 1 to 4, listed in Chapter 2.

The theoretical finding that is important to point out now is the nature of the general equilibrium models. In the epsilon model, just to recall, unemployed and employed workers are homogeneous in human capital endowments. The unemployed have however become the poorest group. Why should the unemployed remain in this situation? They have the necessary human capital to set up firms and produce good B, which could change the equilibrium level of output and the equilibrium degree of inequality.

Several market restrictions impede this course of action. First, there is no market for renting capital. Thus, workers would need to buy physical capital, which in turn needs access to credit and insurance markets. Second, credit and insurance markets are quasi-Walrasian markets. Banks and insurance firms have no incentives to do business with small firms or people that are poorly endowed with physical capital. Banks prefer to supply a small number of large size loans than a large number of small size loans; similarly, insurance firms prefer to supply a small number of large size insurance contracts than a large number of small size insurance contracts. In sum, the workers who are excluded from the labor market are also excluded from credit and insurance markets; thus, even in this model of homogeneous labor, the general equilibrium is with unemployment and inequality. No one has the power or the will to change this situation.

These conclusions also apply to omega and sigma models. The introduction of credit and insurance markets will not alter the general equilibrium result in these societies either. Banks and insurance firms have incentives to exclude the poor. General equilibrium with unemployment, underemployment and inequality will prevail in omega and sigma societies. Again, no one has the power or the will to change this situation.

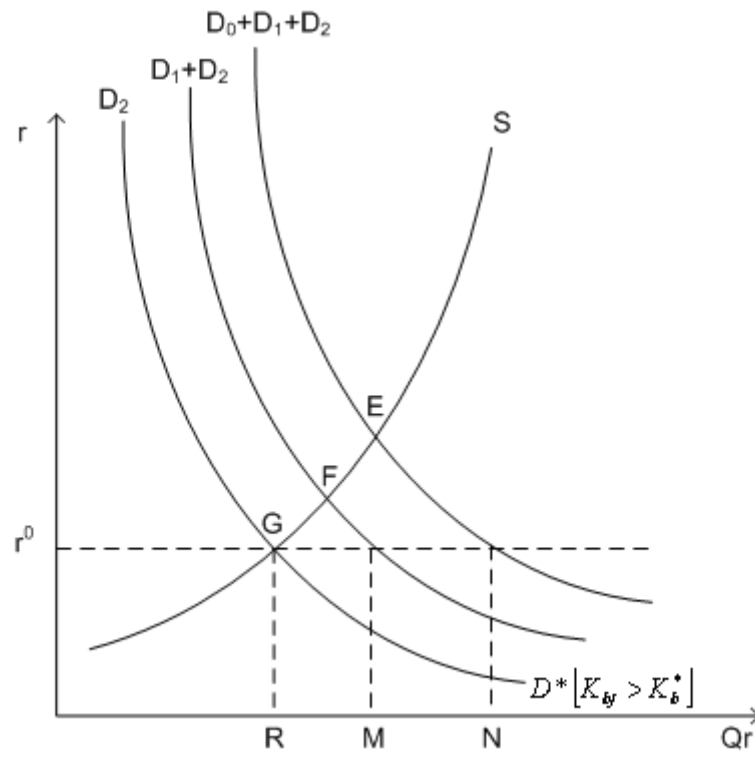
This chapter has shown that labor, credit, and insurance markets play a significant role in the reproduction of inequality in any type of capitalist society. Even under a market structure of perfect competition, markets operate with exclusion mechanisms. In the labor market, the epsilon economy needs unemployment to operate; whereas omega and sigma need underemployment to operate. Workers are unable to become capitalists endogenously due to the nature of the credit and insurance markets. Equilibrium with inequality is indeed an equilibrium situation because no agent has both the power and the will to change it.

A note on market structure is in order. The credit and insurance markets have been analyzed under a market structure of perfect competition. In the real world, we observe that these markets operate with few large firms. Therefore, the assumption of pure competition of the models may seem “unrealistic” and their empirical consistency unconvincing. What needs to be said here is that alternative market structure models of imperfect competition or collective monopoly would show more complex equilibrium solutions. But prices and quantities would still be endogenous. Moreover, it can be shown that, under certain standard conditions, the oligopoly and monopoly models generate reduced form equations that are similar to those obtained under perfect competition; thus, they would still predict the exclusionary nature of these markets. All market structures generate the same beta propositions. Then the conclusions obtained here for the case of perfect competition is applicable to other market structures. The assumption that markets operate *as if* it were under perfect competition is not as restrictive as it appears.

The theory of the market system that has been developed in this book assumes that not all markets are Walrasian. The labor market is non-Walrasian, and credit and insurance markets are quasi-Walrasian. The implication of this assumption is that these three markets play a significant role in the reproduction of inequality and the social structure. Hence, they may be called *basic markets*. Other markets may have an effect as well, but the theory assumes that these effects are not significant. Not all markets are equally important in the generation and persistence of inequality; on the contrary, there exists a hierarchy of markets. When confronted against the real world, the theory of the market system presented here indeed predicts the persistence of inequality in capitalist societies: Fact 7 of the regularities of capitalism, shown in Chapter 2.

The conclusion of this chapter is that the process of physical capital accumulation tends to reproduce inequality and the class structure of a capitalist society. Is the process of human capital accumulation equalizing? This is the theme of the next chapter.

Figure 8.1. Credit Market Equilibrium with Exclusion



CHAPTER 9

EDUCATION AND INVESTMENT IN HUMAN CAPITAL

Human capital is usually defined as the production skills embodied in workers. According to standard economics, human capital plays an important role in the long run output growth of countries. People's skills are as important as machines and technology in the process of economic growth.

Up to now, human capital has remained constant in the models of epsilon, omega, and sigma theories. The aim of this chapter is to develop a theory of investment in human capital that can explain the accumulation of human capital in each type of capitalist society, with special reference to the role played by the initial inequality of society. Standard economics has produced a theory of human capital accumulation, but it assumes a homogeneous capitalist system and thus ignores the role of initial inequality in that process.

A theory of investment in human capital will be presented, with assumptions that depart from the standard theory. The theory will be presented through a particular model. Finally, the predictions of this model will be confronted against the empirical data.

9.1 Static Models with Exogenous Human Capital

What would be the effect of having different endowments of human capital in the static models of epsilon, omega, and sigma shown above?

Firstly consider the epsilon model of Chapter 4, which can be summarized as follows. The human capital distribution among individuals is equalitarian. However, general equilibrium is with inequality. Employed workers receive a uniform market real wage rate because there is one labor market only. Capitalists also receive the homogeneous wage rate, but in addition they received profits. The unemployed received insurance compensation in an amount that is lower than the real wage rate.

Secondly, consider the omega model of Chapter 5. Human capital is also equally distributed among individuals. Again, general equilibrium is with inequality. Wage-earners receive a uniform market real wage rate because there is only one labor market only. Capitalists also receive the homogenous wage rate, but in addition they receive profits. The self-employed have a mean income that is below the wage rate. The unemployed receive no income.

Thirdly, consider the sigma model of Chapter 6. There are two levels of human capital. The high level is equally distributed among capitalists and x-workers, while the low level is equally distributed among z-workers. There is only one labor market, for the high level human capital. X-workers who are wage earners receive a homogenous market real

wage rate. Capitalists receive this wage rate plus profits. X-workers that are self-employed generate a mean income that is smaller than the market real wage rate. Z-workers are all self-employed and generate a mean income that is smaller than the mean income of the x-workers who are self-employed.

The conclusion of these models is that equality of human capital does not imply equality of incomes, neither at the aggregate level, national income, nor for a group of workers endowed with the same level of human capital. The basic reasons are two: (a) capitalists concentrate the ownership of physical capital; (b) workers endowed with the same human capital do not operate in a Walrasian labor market in which full employment equilibrium is reached, but in non-Walrasian markets that operate with excess labor supply.

The higher the endowment of human capital in a capitalist society, the higher the labor productivity level, and the higher the national income level will be. In the epsilon model, the productivity of labor will be higher and thus the labor demand curve will be placed at a higher level. In the omega model, the effect upon total output in the capitalist sector will be as in epsilon; in addition, the productivity of labor in the subsistence sector will also be higher.

The degree of income inequality will not change in any definite direction when the endowment of human capital in a capitalist society is higher. The reason is that total profits might change in any direction. At higher levels of labor productivity, and higher levels of labor demand curve, general equilibrium will be with more wage employment and higher real wage rate; the first effect increases total profits, but the second reduces it; so the net result is ambiguous.

Compare now new endowments of human capital in sigma society with the initial general equilibrium. The characteristics of the initial general equilibrium of national income and income inequality are known. Suppose a new endowment in which only x-workers have more human capital. In this case, the level of national income would be higher and income inequality would tend to increase, as z-workers' income will remain unchanged. Suppose a second case of a new endowment in which z-workers have the same human capital of x-workers, maintaining the total quantity of workers constant. The general equilibrium of production and distribution would now be as in omega society, with one level of human capital. Thus the difference with the initial output equilibrium will be that more workers will be unemployed. The new income inequality may take any direction, because the elimination of z-workers reduces inequality, but the increase in unemployment raises inequality.

The conclusion is clear. The higher the human capital endowment of a capitalist society, the higher will be the level of national income. However, income inequality of a capitalist society will not be lower with higher human capital endowment; it might also be higher or constant.

9.2 Process Analysis of Education

The models shown above assume that human capital is exogenously determined. The need is clear for models in which human capital is endogenous. We need to explain the process of human capital accumulation.

People are not born with human capital. People need to invest in acquiring it through education. Hence, in order to understand the former, we must understand the latter. Education will refer to formal education (school system) only. Other forms of human capital accumulation, such as on the job training, will be ignored. The terms “school” and “student” will refer to each of the levels of formal education (primary, secondary, technical, and university).

Education will be seen as a process that produces general knowledge and human capital. The concept of *process analysis* that was introduced in Chapter 1 will then be applied to education. There are elements that enter into the educational process, elements that come out, and a mechanism that transforms exogenous into endogenous elements, or inputs into outputs.

According to the different disciplines that seek to explain the process of learning (such as psychology, biology, and neuroscience), the endowment of cognitive capacities that students bring to the education process is essential for learning. Humans are born with multiple talents, the so called multiple intelligence theory (Gardner 1999). Although humans are naturally diversified in their talents, the composition is not homogeneous among individuals; thus, some are more talented for art, others for science, and so on. Therefore, individuals are genetically different (not unequal) from each other in their endowments of talents.

Genetically determined cognitive capacities in humans can be considered as exogenously determined (the effect of *nature*). However, these initial capacities or their paths are not given once and for all; they can develop over time in different degrees, depending upon the social environment in which individuals live (the effect of *nurture*). The individual cognitive capacities at each level of education are, in sum, endogenously determined. In the modern literature of neuroscience, this is known as the brain plasticity theory, and it is usually stated as follows:

The brain is not a computer that simply executes predetermined programs. Nor is it a passive gray cabbage, victim to the environment influences that bear upon it. Genes and environment interact to continually change the brain, from the time we are conceived until the moment we die (Ratey 2002, p.17).

The genetic endowments of individual cognitive talents may be assumed to be normally distributed among a given population, the result of a random distribution. But this will not be the rule for the cognitive skills that are developed by the social environment. The important distinction made by Rousseau (1755) refers exactly to these two factors. Rousseau distinguished two types of inequalities among individuals: the *natural*, originated by the gifts of nature, natural endowments, and the random mechanism; and the *artificial*, originated by the functioning of society.

A theory of human capital accumulation will now be presented. A primary assumption of this theory is that the educational process is not uniform for all social classes. Of the two factors that affect the development of cognitive skills of individuals, the essential factor is the social environment, or the social class, to which they belong. The genetic endowments at birth, being randomly distributed, will be considered a less important factor, particularly when we come to aggregate individuals into social classes or groups. Regarding the school system, the assumption is that the quantity and quality of the school inputs is not uniform across social groups.

The alpha proposition of the education process theory can be stated as follows:

α (C).(4)*The theory of Human Capital Accumulation*: In capitalist societies, the educational process is not uniform for all social classes; hence, human capital accumulation takes place under separate and hierarchical educational processes, according to the initial inequality in the individual distribution of economic and political assets within society.

This theory is general. It intends to explain the process of human capital formation in the three types of capitalist societies, epsilon, omega, and sigma.

9.3 The Role of the Initial Inequality

A model of the theory of human capital accumulation can be constructed by introducing three auxiliary assumptions. First, the societies under study will include the three types of capitalist societies together with the social groups that were defined in the models we have been studying so far. Second, students participating in the education process will be endowed with unequal cognitive skills or capacities. Nutrition, health, and early intellectual stimulation are the main channels through which the wealthy can develop higher levels of learning capacity of their children compared to the poor.

Language is another factor of inequality in cognitive skills that is also associated to the socio-economic level of households. There exist language disparities among individuals. This is shown in various aspects of language, such as vocabulary, syntaxes, and ways of speaking, writing and reading skills. In unequal societies, language disparities become language inequality. According to socio-linguistic theory, language inequality is due mostly to social environmental factors than to genetic factors (cf. Hudson 1996, p.204).

When the society is hierarchical, as in sigma society, the language of the dominant social group is also the dominant language. Then language inequality is more significant than language disparity. Moreover, given the segregation that exists in sigma, the command in the use of the dominant language will be unequal between social groups of society; thus the problem of *heteroglosia* will appear as another form of language inequality.

Heteroglosia is a concept that comes from socio-linguistic theory and refers to the existence of various forms or variations in the usage of a given language. This is the case in sigma society, where the usage of the dominant language has a hierarchy, from the one that is considered correct and socially superior (the native language of the dominant social group) to the incorrect and socially inferior one, which is utilized by the subaltern population, and learned as a second language. This is reflected in the problem of different accents and proficiency in the dominant language usage, which will persist even at adult age due to segregation and exclusion (such as “white Spanish”, “mestizo Spanish”, and “indigenous Spanish” spoken in Latin America).

Neuroscience research has shown that we lose flexibility in forming new language connections in our brains by age six or seven. Hence second languages learned after these ages are stored within neural systems that are distinct from those for the native language. By contrast, people who grow up bilingual from birth store their native and second languages in the same neural area (Ratey 2002, p. 278).

In sigma society, language inequality will remain even at the end of the educational process. Language then becomes a social marker: “Let me hear how you speak and I will tell who you are”. This social marker refers of course to the dominant language, which is a second language for z-populations. There will be path dependence on the language inequality process.

Inequalities in the language skill among social groups will imply unequal cognitive skills of their children. There are at least two reasons for this: the heteroglosic effect and the oral language effect. The first was presented above and has to do with the unequal command on the dominant language, a result of segregation. On the second, the theory is that abstract and complex thoughts are not only language-dependent, but also complex language-dependent. Philosopher John Searle has stated this theory as follows:

“Some thoughts are of such complexity that it would be empirically impossible to think them without being in possession of symbols. Mathematical thoughts, for example, require a system of symbols. ...Complex abstract thoughts require words and symbols” (Searle, 1995, p. 64).

The implication of this theory is that written language allows people to work with more abstract and complex thoughts than does oral language alone. Therefore, oral language societies will show disadvantages in cognitive skills compared to written language societies.

Consider as context an epsilon society in which people live in a written society and most of them are literate. By contrast, consider a sigma society, a multilingual, multicultural, and hierarchical society, in which x-workers live in a written culture and most of them are literate, but z-workers live in an oral social environment (because the aboriginal language is not a written language) and most of them are illiterate in the dominant language. Inequality in language skills will be higher in sigma compared to epsilon.

In the case of sigma society, z-populations’ children will be limited in the learning of abstract and complex thoughts because they come from an illiterate environment. This is the first effect of language on cognitive skills, the illiteracy effect. But, because they come from an oral culture as well, where the aboriginal language is not written, those limitations will be reinforced. This is the second effect, the oral culture effect. Hence, z-populations’ children will have, on average, a lower cognitive skills than x-populations’.

The model also implies that inequalities in learning capacities will exist even between illiterate families, depending on whether they live in epsilon society or sigma society. Students that come from an illiterate social environment within a written culture will be less handicapped in learning skills compared to those coming from an oral culture.

The inequality in the endowments of economic and political endowments among individuals, therefore, leads to linguistic inequality, which in turn leads to inequality in cognitive capacities of students. A hypothesis coming from socio-linguistic theory goes even further: “Linguistic inequality can be seen as a *cause* of social inequality, but also as a *consequence* of it, because language is one of the most important means by which social inequality is perpetuated from generation to generation” (Hudson 1996, p.205).

In sum, the model of the theory of human capital accumulation assumes that in capitalist societies, in which household are endowed with unequal economic and political assets, students participating in the educational process will be endowed with unequal learning capacities. This is part of the mechanisms underlying the educational process because learning capacity is unobservable.

9.4 Transforming Education into Human Capital

On the educational process, seen as a process that produces general knowledge and human capital, the model assumes that the school inputs are the exogenous variables, the general knowledge and human capital the endogenous variables, whereas learning capacities of students and the educational technology are the mechanisms that transform the inputs into outputs. The model also assumes that the learning capacities of students are not homogeneous, as shown above.

Consider for a moment that schools are all homogeneous. Even under this condition, for given years of education, the children of the wealthy will accumulate more human capital than the children of the poor, due to the initial inequality in learning capacities. But children will not have the same years of schooling. The accumulation of human capital requires financing. Rich households have greater financing capacity than poor households, which allows them to finance more years of schooling. Then, the income effect on investing in human capital is positive: the quantity of human capital demanded will positively depend upon the level of household income. Consequently, on average, the children of rich households will have a higher level of human capital than the children of the poor households on two accounts: more years of schooling and higher learning capacity.

If the assumption that schools are homogenous is now abandoned, differences in the school quality will be another factor of differentiation. The effect of such difference is that the children of rich households assist to private schools, while the children of poor will attend public schools. Private schools are highly equipped with inputs and technology and have better trained teachers than public schools.

This simple model predicts that the children of rich households will acquire, on average, not only more schooling years but also a higher level of human capital for every year of schooling than the children of poor households. The transformation of education into human capital will operate differently for different social groups.

It should be clear that this theory of investment in human capital assumes that schooling years is not the same as human capital. The general relation between schooling years and human capital levels produced is positive. However, this relation will not be unique, but it will take different forms depending on the type of capitalist society.

The model can be summarized in the following set of equations for sigma society:

$$k_h(t+1) = f^j(E(t), S_j), f_i > 0, \text{ and } j=Z, X, A \quad (9.1)$$

$$E(t) = g^j(y(t), S_j), g_i > 0 \quad (9.2)$$

$$k_h(t+1) = F^j(y(t), S_j), F_i > 0 \quad (9.3)$$

The first equation connects education or schooling years (E) and human capital (k_h) for each social group S (including here groups Z , X , A). Time t refers to periods or generations. The human capital of today is the result of education of the previous period. The second equation says that years of schooling depend positively on the income level (y) of the social group. The third equation is just the reduced form of the two previous structural equations; it indicates that the human capital of the following period depends ultimately upon the income level of the social group in the current period. The effect of the social group is in all cases positive. Because the equations refer to social groups, the variables should be read as the mean values of the corresponding social group.

Equation (9.1) is represented in Figure 9.1. Consider the sigma society, in which the social structure is composed by three hierarchical socio-economic groups: Z , X and A . In this society, there will be a positive relation between years of education and levels of human capital accumulation for each social group; however, this relationship is *separate* for each social group, and also *hierarchical* between the groups. The implication is that if the schooling years were the same, the human capital accumulated would be the highest for the social group A and the lowest for the group Z , with the group X in between. But the years of education between social groups are not the same; the highest is for group A , the lowest for group Z , with group X in between.

The transformation of education into human capital, therefore, travels along separate and hierarchical trajectories, indicating the differences in both the quality of students and the quality of the schools. The z -population will have access to public schools of the lowest quality because they are second rate citizens. Therefore, the difference in trajectories also reflects the particular form of functioning of democracy in the sigma society. Citizens of different categories have access to local public goods of different categories. Thus the transformation of education into human capital does not take the same form, along the same trajectory, for all social groups.

In epsilon and omega societies, which are socially homogeneous societies, the transformation from education into human capital will operate along two trajectories only. In these societies, the z -population does not exist. Therefore, Figure 9.1 will now show two curves, one for social group A and the other for group X , but still separate and hierarchical, reflecting the social class structure of the society.

Equation (9.3) is represented in Figure 9.2. In sigma society, there are three social groups Z , X , and A . For each social group, higher mean income of the households implies higher human capital for their children. However, the differences between social groups imply that the positive curve relating income to human capital is placed at different levels and at different income ranges. In the case of epsilon and omega societies, there will be just two curves, X and A .

These relationships imply that, in any capitalist society, the educational process is not conducive to the equalization in the human capital of social groups. The poor and the rich accumulate human capital along different paths. The educational process may reduce the initial inequality in schooling years between social groups, but it will not equalize human capital. The same years of education by social groups will not imply the same level of human capital. The school system thus tends to reproduce the initial inequality in every type of capitalist society.

These models predict that in every type of capitalist society (epsilon, omega, or sigma), children will inherit the *relative* position of their parents in the human capital distribution. In the process of human capital accumulation, the initial human capital gaps between social groups are not reduced endogenously. Because there exist ceilings to education years, inequality in education years between social groups may become reduced endogenously, but human capital will not. Even if education years could become equalized endogenously, the human capital of the different social groups will not become equalized. Thus, the educational process is not human capital equalizing. In the process of human capital accumulation, there is a path of dependency; that is, initial conditions matter.

The British biologist Francis Galton (1869) established long time ago a positive correlation between the heights of children and those of their biological parents; moreover, he found that the height differences among the children were smaller compared to the differences among parents, which he called the “law of regression towards the mean” in the biological process between generations. Height inequalities tend to diminish.

The dynamic models of the human capital accumulation theory presented here predict that, in the education process, the “regression towards the mean” could hardly occur in human capital inequality. In inter-generational terms, “children” of a social group will tend to inherit the relative economic position of their “parents” of that particular social group. There will be educational mobility, but not human capital mobility.

9.5 Exclusion vs. Wage Discrimination in Labor Markets

If human capital and schooling years are not equivalent, as shown by this theory, do firms buy education or human capital in the labor market? Consistent with their rationality of profit maximization, firms will buy human capital in the labor market, which is the relevant factor for productivity and profits. Each level of human capital will then constitute a particular labor market. Furthermore, in a competitive labor market, wage rates will be uniform for the same level of human capital, not for the same level of education.

If differences in wage rates per schooling years were observed empirically, would this fact refute the model? No. The observation that firms pay different average wage rates to workers with equal years of schooling does not constitute a case of wage discrimination, as the literature of standard economics calls it. According to our theoretical models, wage discrimination exists if, and only if, firms pay different average wage rates to workers with equal human capital.

Why might wage discrimination by human capital exist? Consider a sigma society in which x-workers and z-workers have similar levels of human capital, but z-workers are paid less. The origin of this discrimination could come from productivity factors other than human capital, such as ethnicity. For example, preferences of consumers for some goods or services could depend on whether it is produced by z-workers or x-workers. There would be labor market segmentation: the derived demand for z-workers would be different compared to the one for x-workers.

It could also be originated in a problem of incomplete information in the labor market. In the short run, capitalists could have little confidence on z-workers because of cultural differences and ethnic prejudices (which will tend to disappear in the long run). In

this case, transaction costs of employing z-workers would be higher in relation to x-workers. As a consequence, labor market segmentation would again appear in the labor market: given equal level of human capital, z-workers would obtain in the labor market an inferior wage rate compared to x-workers to compensate the higher transaction costs.

The empirical prediction of the labor market mode in sigma society is that the low wages of z-workers relative to those of x-workers are generated mainly by exclusion effects rather than by the wage discrimination effect. Exclusion takes two forms: exclusion from good quality education (less human capital per equal years of education) and exclusion from the quantity of education (less years of schooling). Profit maximizing firms have incentives to pay the same wages for equal human capital level, which rules out discrimination, except for the minor situations shown above.

Exclusion, not discrimination, is thus the essential factor explaining total labor income (wage and self-employment income) differences between z-workers and x-workers. For example, according to this model, the income gap between group Z and group X cannot come from the fact that engineers of group Z receive a smaller wage rate compared to engineers of group Y; the major factors would include the effect that the proportion of z-workers that are engineers is lower than that of x-workers, and the effect that the proportion of engineers endowed with a high level of human capital is lower in group Z compared to group X.

Regarding omega and epsilon models, the results will be different. Because z-workers do not exist, there will be less inequality in human capital among workers and thus the degree of inequality in real wages will be lower than in a sigma economy.

9.6 Empirical Predictions

It would be useful to show that the situation shown in Figure 9.2 constitutes in fact an equilibrium situation; namely, that there is no social actor that has both the power and the will to change this situation.

Consider the sigma society. In this society, education is transformed into human capital along different paths for different social groups, as shown in Figure 9.1. The basic markets (labor, credit, and insurance) operate with exclusions. In particular, the credit market could not finance the accumulation of human capital of the x-population, much less that of the z-population. On the other hand, the access to basic public goods (education and health) is also differentiated by social groups. This is how markets and democracy, the two basic institutions of capitalism, operate in sigma society. In epsilon and omega, the result is qualitatively the same.

When it comes to education, public policies seem to be directed to creating equal opportunity for all. High rates of enrollment and expansion of public schools would indicate that education is an income equalizing system. This model predicts that this is not the case. Figure 9.1 allows us to define analytically the content of the *equal opportunity principle* in education—a single human capital path for everyone. In a sigma society, for example, equal opportunity policy means the three curves A, X, Z should be reduced to only one; that is, it implies the shifting of the X and Z curves onto curve A. Only then, the same years of education would imply the same human capital level for everyone.

Governments do not have the incentives to carry out such policies in sigma societies. Z-workers are second rate citizens and thus have no political voice to push governments towards this policy. Governments act guided by the motivation of maximizing votes in the next elections, that is, in the short run. Accordingly, governments will seek to supply workers' children with more years of education, inaugurating more school buildings, which is politically more profitable in the short run. This policy will imply maintaining the separate trajectories for different social group.

Could workers get equal education opportunity policies through collective action? Collective action is subject to several constraints. The Olsonian problem of the free rider is one of them (Olson 1965). In particular, z-workers are too poor to finance collective actions for such complex set of development policies. Finally, z-workers are socially excluded; they are second-rate citizens, with no political voice, which makes collective action less likely to succeed. Therefore, the equilibrium shown in Figure 9.1 is indeed, the equilibrium situation in sigma society.

In the cases of epsilon and omega societies, governments and the people will interact more to implement policies of equal opportunity. The democratic system is more participatory and basic markets are less restrictive. This effect goes in the direction of creating equal opportunity policies. But the initial inequalities in the endowments of economic assets will not allow the curve Y to get shifted onto the curve A. Then workers will advance in years of education, but along the curve Y. This is too, the equilibrium situation.

An exogenous variable of the model is the initial inequality in asset endowments among individuals (δ). As long as inequality in the distribution of assets remains constant, the differentiated paths will also remain unchanged. If this initial inequality were reduced, then the effect of the socio-economic background of the student would also be reduced; the education would still follow different paths, but along more equalizing paths between social groups. Consider the case of sigma society. A redistribution of economic and political assets would imply a change in the functioning of democracy and higher political voice for z-workers, which would produce better quality of public education, together with better quality public of health services. Thus, the same years of education would be transformed into higher levels of human capital for z-workers, reducing the gaps with the social groups X and A. In Figure 9.1 the line Z will be shifted upwards closing the gaps with the lines X and A.

9.7 Empirical Consistency: The Capitalist World

The empirical predictions derived from the model of the theory of human capital accumulation can be confronted now against the results of the international empirical literature. In general, the empirical studies on these relations are scarce due to the unavailability of human capital data.

The model predicts that schooling years varies by social groups—Figure (9.1)—in each capitalist society. A study on Peru, a sigma society, which is based on the national household survey of 2003, found that indeed education levels by social groups were statistically significant. The indigenous population represented social group Z, and the mestizo represented social group X, and the white upper and middle classes represented the

social group A. The mean years of schooling among adult populations were 6.9, 10.8, and 13.7 (Figueroa 2010, Table 2, p.121).

The model also predicts that the level of human capital depends upon the income level of social groups in a given society—Figure 9.2. The surveys of PISA (Program for International Student Assessment) provide a relevant database for international comparison. This survey measures learning levels at school by the performance of 15-year-old students in reading, mathematics, and science tests applied in the OECD countries and other participating countries. The 2009 results (OECD 2010) show the big gap between the First World and the Third World: from the 16 countries that belong to the Third World, only South Korea and Singapore show average marks that are comparable to those of the First World; these marks were well below in the other 14 countries. The report also states: “Socio-economic background of students and schools does appear to have a powerful influence on performance” (p. 5).

The last result is certainly consistent with the model. The mix of A-Y-Z differs between groups of countries. Z-populations are predominant in the Third World with strong colonial legacy, but not in the Third World with weak colonial legacy (South Korea and Singapore).

Biologist Jared Diamond (1999), in trying to explain why the Spaniards conquered the Incas and Aztecs, and not the other way around, has included written language as one of the explanatory factors for Western superiority. He says, “Writing marched together with weapons, microbes, and centralized political organizations as a modern agent of conquest” (p.216). This argument is in accord with the prediction of the model, which assumes that language inequality plays a crucial role in the human capital accumulation in sigma societies.

9.8 Conclusions

The available empirical facts do not refute the predictions of the model of the theory of human capital accumulation presented here. Therefore, there is no reason to reject the theory at this stage of our research. The education process in capitalist societies is not human capital equalizing. Human capital accumulation follows different paths depending on the socio-economic background of individuals.

Access to credit and insurance markets by the poor could make, through financing education and protecting human capital, the education process more favorable to human capital equalization. However, as shown in chapter 8, these markets exclude the poor. Hence, credit and insurance markets indeed constitute basic markets, as they play a significant role in reproducing inequality in capitalist societies.

If education were human capital equalizing, the transformation of education into human capital would have to follow a unique path, independent of the socio-economic background of people. In terms of Figure 9.1, it means that the two lines Z and X should be shifted unto the line A; hence, the same years of schooling would generate the same human capital. This is a necessary condition, but not a sufficient one for human capital equalization. In terms of Figure 9.2, equalization in the number of years of schooling would also be needed. Social group Z, with income level y_1 , can reach only the level M of human

capital, whereas social group X, with income level y_2 , can reach level N; equalization in human capital would require that these two social groups could reach level R, which corresponds to social group A, with income level y_3 , in spite of the current income gaps.

Conceptually, this could be a definition of *equality of opportunity* in the educational process. Certainly, this is not what happens in the real world. Particularly in the Third World, governments publicize high rates of school enrolment among the poor, but hide the fact that these children are travelling along different paths compared to the children of the wealthy households. The rationality of maximization of votes for the next elections could hardly lead governments to apply the equal opportunity policies. Facts have revealed their preferences: governments just do not have the incentives to do it.

The educational process produces two types of outputs: human capital and also general knowledge, including here social norms. The first output reproduces inequality in human capital in the First World and in the Third World, as shown in this chapter. In addition, the school system in the Third World seems to reproduce the social norms of citizenship inequality as well.

Up to now, this book has shown that epsilon theory seems to explain the process of production and distribution in the First World countries, omega theory of that in the Third World countries that have weak colonial legacy, and sigma theory of that in the Third World countries that have strong legacy of colonialism. These explanations correspond to each type of capitalism, taken separately. As partial theories of different types of capitalist economies, these theories are not refuted by the known facts of the real world. Thus there is no reason to reject these partial theories at the present stage of our investigation.

The remaining question is whether these valid partial theories can generate a valid unified theory of capitalism, which may be able to explain production and distribution in the capitalist world, taken as a whole. Valid partial theories do not necessarily generate a valid unified theory, as we know from the experience of physics. The construction of the unified theory of capitalism is the theme of the next chapters.

Figure 9.1. Theoretical Relations between Education and Human Capital, by Social Groups

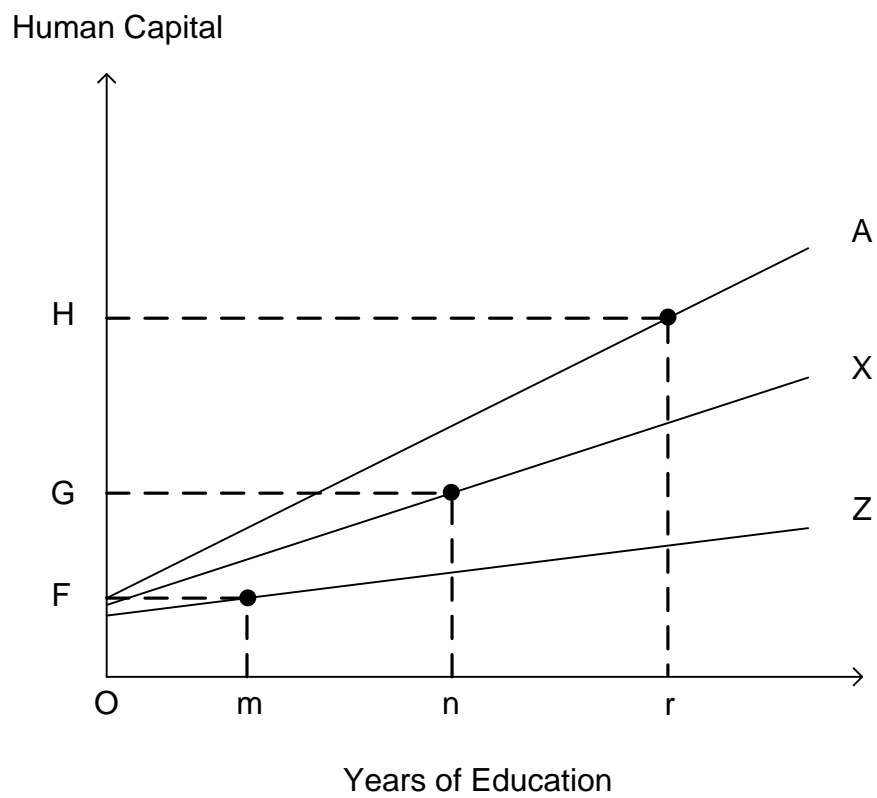
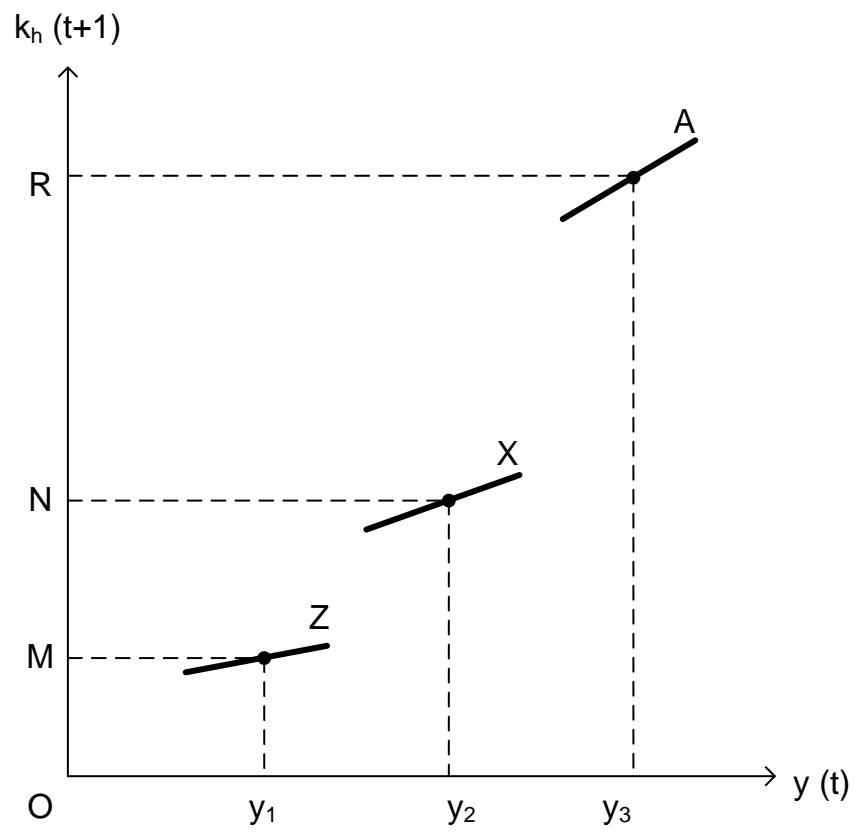


Figure 9.2. Theoretical Relations between Income and Human Capital, by Social Groups



PART III

A UNIFIED THEORY OF CAPITALISM

CHAPTER 10

PRODUCTION AND DISTRIBUTION IN THE CAPITALIST SYSTEM

The principal aim of this book is to understand the development process of the capitalism system: output growth and distribution. For that purpose, three abstract capitalist societies have been presented in this study: epsilon, omega and sigma. As shown in Chapters 4 to 6, these abstract societies do in fact explain the basic features of the First World and Third World taken separately.

Can a unified theory of capitalism that is consistent with those three separate theories be constructed? If such unified theory were found, we could explain the functioning of the capitalist system taken as a whole.

Good partial theories, however, do not necessarily imply a good unified theory. Consider, for instance, general relative theory and quantum theory in physics. The former explains the large physical world, whereas the latter explain the small subatomic physical world. Thus they explain these two physical worlds, taken separately, but they cannot explain the physical world taken as a whole. These theories are contradictory to each other, that is, they both cannot be true; hence, a unified theory of physics—the *theory of everything*—is the fundamental problem being researched on today (Hawking 1996).

Similar challenge appears in this book now. Chapters 7 to 9 have develop some relationships that are intended to make the transition from partial theories to the unified theory. This chapter starts the construction of the unified theory.

10.1 Foundations of the Unified Theory

In the construction of a unified theory, we will follow the *Moore principle*, named after E. H. Moore, an American mathematician. This principle says that the existence of analogies between central features of partial theories leads to the existence of a unified theory, which will represent the underlying central features of those partial theories (cited in Samuelson, 1965, p. 3).

In order to become a unified theory, therefore, the partial theories must give rise to common central features, as Moore's Principle requires. In terms of the alpha-beta method, the "central features" correspond to alpha propositions; hence, the requirement is equivalent to finding a common set of alpha propositions in epsilon, omega, and sigma theories. This common set, if it exists, would then constitute the alpha propositions of the unified theory; if it did not, there would not be unified theory. In a way, the requirement is that there exists a set of alpha propositions that are not contradictory to each other.

The common primary assumptions of epsilon, omega, and sigma societies (presented in chapters 4, 5 and 6) are easily identifiable. The corresponding unified theory can then be expressed as the following set of alpha propositions:

$\alpha(c)$ (1). *Institutional Context*: (a) Rules: People participating in the economic process are endowed with economic and political assets; economic assets are subject to private property rights; people exchange goods subject to the norms of market exchange, which include the norm that nominal wages cannot fall; the market system operates with Walrasian, quasi-Walrasian, and non-Walrasian markets, in which the labor market is of the latter type. The political regime is democratic. (b) Organizations include firms, households, and the government.

$\alpha(c)$ (2). *Initial Conditions*: There are different types of capitalist societies, which differ by two initial conditions: factor endowments and the initial inequality in the distribution of economic and political assets among individuals. Initial inequality makes capitalist societies class societies, constituted by capitalists and workers.

$\alpha(c)$ (3). *Economic Rationality of Agents*: People act guided by the motivation of self-interest. Capitalists seek two objectives that are hierarchically ordered: maintenance of the social position and profit maximization. In the labor market, workers seek to maximize wages and minimize effort. The class division of society generates social conflict on income distribution and work effort; for the latter, capitalists use devices to extract work effort from workers. Politicians also act guided by the motivation of self-interest.

$\alpha(c)$ (4). *Social Tolerance to Inequality*: Individuals have a sense of justice or fairness with respect to economic inequality, which implies a limited tolerance to inequality. Whenever inequality reaches a level beyond the tolerance threshold, individuals will protest and seek to restore inequality to the tolerable values.

This set of alpha proposition constitutes the common alpha proposition of the partial theories epsilon, omega and sigma; moreover, this set is a logical system, as there are no logical inconsistencies between them. Therefore, this set constitutes the alpha propositions of the unified theory of the capitalist system. The symbol “c” stands for capitalism as a whole. The three abstract capitalist societies that have been developed in this book now become partial theories of this general theory; that is, there is no logical internal contradiction between the primary assumptions of the unified theory and those of the partial theories.

It will be shown in this chapter that indeed the three theories constitute partial theories of the unified general equilibrium theory. We have seen in each theory how a general equilibrium model was constructed as the aggregation of partial theories of markets, in which the interactions between markets were taken into account; the same principle will now be applied in the unified theory, in which the three types of societies, as partial theories, will be aggregated and its interactions taken into account. Unity of knowledge in the unified theory will thus be assured.

The unified theory seeks to explain the empirical differences between the First World and the Third World, which were shown in Chapter 2 as Facts 6 and 7. Static and dynamic models will be constructed for that purpose. The static model intends to explain

these differences in a given period, which implies that the capital stock and technology are given in each society. This static model will seek to explain the *income level* differences between capitalist countries and will be presented in this chapter.

10.2 A Static Model: Explaining Income Level Differences

The question of why the First World countries are richer than the Third World countries, as Facts 6 indicates, will be answered with a static model of the unified theory.

Chapter 7 presented static models for each of the three types of capitalist societies, taken separately. As shown in equation (7.5), in each society, the endogenous variable was national income and the exogenous variables included international prices (terms of trade and interest rate), factor endowments, technology, and the initial inequality; moreover, the effect of changes in international prices upon total output was to generate short run fluctuations around the equilibrium *level* of output, which was determined by factor endowments, technology, and the initial inequality.

Transforming those three partial models into a unified model implies the introduction of interactions between the three societies in its integration into a general equilibrium model. In the unified model, a general equilibrium model of the capitalist system, some variables that were exogenous for each society will now become endogenous and some will remain exogenous. In the short run, the interactions between the societies are reduced to market exchange of goods, which implies that the international prices will become endogenous in the unified model. However, as just stated above, changes in international prices will cause short run fluctuations around the *level* of total output, but not in the level itself. Because we are interested in the level of output, in the construction of a long run static model, the effect of changes in the international prices may be ignored.

The static model of the unified theory presented here will seek to explain the differences in the level of total output among the three societies that constitute the capitalist system. To make the comparison meaningful across societies with different population sizes, output per worker will be utilized as the endogenous variable, instead of total output. Hence, the endogenous variables of the model will be output per worker, whereas the exogenous variables will include factor endowments, initial inequality, and the level of technology.

The integration of the three static models into a static model of the unified theory requires some auxiliary assumptions. They are:

- (a) The effect of changes in international prices (terms of international trade and international interest rates) will be ignored.
- (b) The three types of societies will produce a single good. By assuming the relative prices of these goods constant, we can invoke the well-known Hicksian Composite Good Theorem, and treat them as just one composite good.
- (c) The level of technology (A) utilized in the capitalist firms is the same across societies and is subject to constant returns to scale. The implication of this

assumption is that, given the level of technology, double quantity of inputs will produced double quantity of total output; hence, output per worker (y) will depend only upon capital per worker (k). “Capital” will also be a composite good, which includes physical and human capital (ignore infrastructure for the sake of simplicity).

According to these assumptions, equation (7.5) can then be re-written in terms of output per worker (y) as follows:

$$y^0 = f^j(A, k_j, \delta_j), j=\varepsilon, \omega, \sigma \quad (10.1)$$

$$f_1 > 0, f_2 > 0, f_3 < 0$$

$$k_\varepsilon > k_\omega > k_\sigma$$

$$\delta_\varepsilon < \delta_\omega < \delta_\sigma$$

The equilibrium level of output per worker (or national income per worker) is determined in each society j by the level of technology, factor endowments, and initial inequality. As shown in the partial theories, for given factor endowment or capital per worker (k), total output will be the largest in each society when equilibrium takes place at effective full employment in the capitalist sector. Although static equilibrium need not reach this situation, we can assume that this will be the case, just to make the comparisons meaningful across societies. Therefore, in what follows for each capital per worker endowment there will be a particular *level* of output per worker, which corresponds to the value of the static equilibrium under effective full employment in the capitalist sector.

Equation (10.1) indicates that societies differ not only in the function, but also in the range of values of the exogenous variables. However, the effects of the exogenous variables are similar in all societies. The effects of technology and factor endowments are positive. The effect of the initial inequality is negative because it leads to higher social disorder, which affects negatively labor productivity in the capitalist sector.

This model is now represented in Figure 10.1. The horizontal axis measures capital per worker and the vertical axis output per worker. The line OR shows the relationship between these variables for epsilon society, for given level of technology and initial inequality: the capital per worker endowment determines the current output per worker, marked at point c.

Omega society is endowed with lower capital per worker, which could imply an output per worker at point c' , along curve OR, if omega operated with a capitalist sector only. But omega's capital per worker endowment is k_ω , which is below k^* (the threshold of overpopulation, where marginal productivity of total labor is zero); hence a subsistence sector exists in which output per worker is smaller than in the capitalist sector, which implies a lower aggregate output per worker, market at point b, along the lower labor productivity curve OP. (Just for the sake of simplicity, ignore the effect of the initial inequality differences with epsilon society). If the capital per worker endowment increased exogenously to k^* , then omega society would produce at point c^* ; if this endowment increased even further, omega would produce along the curve c^*R , which corresponds to the productivity curve of epsilon.

Sigma society's capital per worker endowment is k_σ , the lowest value in the group. If this were the only difference with epsilon, sigma could product at point c". However, sigma has a subsistence sector in which x-workers are self-employed, which implies that the aggregate output per worker will fall and could reach only point b', along the productivity curve OP of omega; moreover, there exists the subsistence sector in which z-workers are self-employed, which implies a further drop in output per worker, and thus output per worker reaches point e', along the labor productivity curve ON. Finally, there is the negative effect of having a higher initial inequality, which implies another downward shift of the productivity curve to the level OM, and the equilibrium output per worker reaches only point e.

Given the differences in factor endowments and the initial inequality, three levels of labor productivity curves are then found, one for each type of society: OR corresponds to epsilon, OP to omega, and OM to sigma. There is an equilibrium output per worker in each society, which reflects output per worker in the capitalist sector in epsilon; a weighted average of the output per worker in the capitalist sector and in the subsistence sector in omega; and a weighted average of output per worker in the capitalist sector and in the two subsistence sectors in sigma. The terms s_x and s_z in the graph indicate the share of total labor in the subsistence sectors.

The equilibrium conditions of the unified static model presented in Figure 10.1 can explain the observed income level differences between the First World and the Third World as follows. The income gap between epsilon and sigma, the distance between point c and point e, is due to differences in the levels of labor productivity, which in turn are explained by the differences in two exogenous variables: factor endowments and initial inequality. The gap between epsilon and omega, the distance between point c and point b, is just due to the effect of differences in factor endowments. The empirical differences in income levels in the world capitalism in two points of time, 1980 and 2008, which were shown in Table 2.2, are thus given a scientific explanation.

The unified static model predicts that exogenous changes in the ratio capital per worker will affect output per worker in each society along its particular curve. Although the three societies have the same technology level, the productivity curves are separate and hierarchical. It would be a mistake to take the curve F, which connects the equilibrium points of each society, as the uniform production function for the three societies. The implication of this analytical distinction is that from observed values of output per worker and capital per worker in the First World and the Third World we cannot infer that if the latter had the same capital per worker of the former, output per worker would be equalized. This equalization could occur between the First World and the Third World with weak or no colonial legacy only, which is just a small part of the Third World.

According to the unified static model, income level equalization between the First World and the Third World would require not only capital per worker equalization, but also initial inequality equalization; more analytically, it would require the transformation of the Third World from sigma society into omega society. The Third World would have to go through qualitative transformations, not only through quantitative changes.

10.3 Within-country Inequalities

Now, let's turn to inequality differences within countries. Equation (7.6), in Chapter 7 above, showed that the level of income inequality in the long run is determined by the initial inequality in asset distribution in each society, with short run variations around this value, given by the effect of changes in the other exogenous variables.

By assumption, the degree of the initial inequality in the asset endowments among individuals is the highest in sigma and the lowest in epsilon, with that of omega in between. Assets include physical capital, human capital, and degree of citizenship. Therefore, the static model predicts that the degree of income inequality is the highest in sigma society and the lowest in epsilon society, whereas the value for omega lies in between.

It should be noted that the assumption of differences in the initial inequality is consistent with the other about differences in factor endowments. Given the size of the capitalist class (who concentrate the property of physical capital), more population of workers will certainly increase inequality in the distribution of physical capital; so overpopulated societies will show a higher degree of inequality in the distribution of physical capital. On the other hand, the distribution of human capital in two overpopulated societies, omega and sigma societies, assumes that in sigma society z-workers will have lower human capital endowment than x-workers; consequently, sigma society will show a higher concentration of human capital compared to omega society, in which only x-workers exist. Hence, societies with low factor endowments also tend to show higher degrees of initial inequality.

Hence, equation (7.6) can be re-written as follows:

$$D^0 = G(\delta), G' > 0 \quad (10.2)$$

The degree of inequality in the distribution of the *stocks* of resources or assets determines, in the static equilibrium under effective full employment in the capitalist sector, the level of the degree of inequality in the distribution of the *flow* of income. Because sigma society has the highest initial inequality, it will also have the higher degree of income inequality. Epsilon society has the lowest initial inequality and also the lowest income inequality. Omega situation will lie in between.

The static model therefore predicts that Third World countries with significant colonial legacy will show on average a higher degree of inequality, compared to those with weak legacy. First World countries will show on average the lowest degree of income inequality. This is consistent with the Gini index data shown in Table 2.2, Chapter 2, in which the differences in inequality between these groups of countries in each of the two periods are remarkable.

10.4 The Question of Short Run Changes

Differences in the levels of output per worker and income inequality between capitalist societies are, according to these results, caused by structural factors only: technology, capital per worker, and the initial inequality. What factors do explain short run variations of output per worker and income inequality?

International prices were exogenous in the partial theories. Changes in international prices caused variations of output and distribution around the levels of equilibrium in the static model. In the static unified model, international prices are endogenous; hence, these variables cannot be causal factors. However, because some countries are price-makers, due to their economic power, and are able to set international prices, such as the nominal interest rate by the US Federal Reserve and the oil nominal price by OPEP, some international prices may remain as exogenously determined in the model.

But would these social actors like to change these prices arbitrarily, independent of the economic process, that is, exogenously?

This problem is similar to that of the monopolist, who is a price-maker, but who cannot set the market price arbitrarily (exogenously). The monopolist seeks to maximize profits and seeks to set the market price accordingly; hence, the monopolist cannot choose any arbitrary price, but the particular price that makes profits the largest. If the demand conditions change, then the monopolist will adjust the price according to the profit objective. The monopolist is a prisoner of his selfish motivation. The monopoly price is thus endogenous.

Similarly, this static model of the unified theory assumes that social actors that are price-makers will set the international prices according to their own interest and will change them when the situation changes, that is, the model assumes that international prices are endogenous. Their variations are not the cause but the consequence of the economic process. These prices may be a *proximate* factor of changes in production and distribution in the capitalist system, but not an *ultimate* factor.

Models presented in macroeconomic textbooks assume that fiscal and monetary policies are exogenous, that is, governments can choose policies arbitrarily. In terms of monetary policies, these textbooks show that central banks have three policy instruments to choose from: nominal interest rate, money supply, and nominal exchange rate; moreover, there is only one degree of freedom: once one instrument is chosen, the other two will be endogenously determined. This may be true for one country taken separately (in partial equilibrium models); in general equilibrium models, when there are interactions between countries, and between central banks, our model assumes that all these nominal variables will become endogenous. Moreover, the static models by types of societies presented above (Chapters 4, 5, and 6) showed that monetary and fiscal policies are not truly exogenous, but quasi-exogenous; hence, they will tend to be endogenous in the aggregate of societies, in the static model of the unified theory.

Expectations (about changes in market prices and quantities) also appear as an exogenous variable in most macroeconomic textbooks. This is ruled out in this model because, in the aggregate, it could hardly change independently of the economic process; but more importantly, a model that includes expectations, which is unobservable, will become empirically unfalsifiable.

What will then enter into the static model of the unified theory as short run exogenous variables? They are the factors that are independent of the global economic process, such as natural disaster shocks. Debt crisis could also be considered a kind of shock. Foreign debt of a country is an endogenous variable (which has been ignored in the model); however, as the debt accumulates beyond a threshold value, the mechanical economic process is transformed into an evolutionary one, because qualitative changes will

appear in this process, which will end up in a debt crisis. The debt crisis then becomes a shock.

These shocks (natural disasters and debt crises) can then operate as changes in exogenous variables in the short run, which will lead to changes in nominal prices and then to changes upon production and distribution, not only in the indebted country, but also in the capitalist system as a whole. These shocks will have a short run effect, as they will cause variations in the equilibrium values of the output per worker and income inequality in society. But those variations will occur around their values of static equilibrium, which is determined by structural factors.

10.5 The Question of International Trade

The static model of the unified theory assumes that capitalist societies are open to international trade. But trade does not play any role in the determination of the level of output per worker or in the level of inequality. It does not play any role in the determination of the level of real wages either. These levels are determined once factor endowments, technology, and initial inequality are given.

However, in the short run (for given factor endowments and technology), international trade does have an effect upon production and distribution in each type of capitalist society. As shown in the partial theories (Chapter 4 to 6), changes in the international terms of trade will modify the equilibrium values of the endogenous variables in a capitalist society, such that the variations of the endogenous variables will occur around their level values. Thus changes in the international terms of trade will cause changes in the real wage rate and labor productivity (output per worker), but only around their level values, within a limited range of possible values.

What are the factors that determine the pattern of trade? In the short run, the comparative cost advantage of countries in the production of goods will depend upon the differences in the levels of labor productivity and real wage rates. Consider a world of two goods only, say B and C. Suppose epsilon societies have a higher level of labor productivity in both goods relative to sigma societies, but the difference is larger for good B. Epsilon societies will produce good B and sigma societies good C.

What about differences in real wage rates? What if the level of real wage rate is higher in sigma than in epsilon? If this were the case, no comparative advantage would exist. But, the level of real wages is not independent of the level of labor productivity. The reason is very simple: profits of firms come from the difference between average productivity of labor and marginal productivity of labor, which is equal to the real wage rate. This important relation can be illustrated in Figure 10.1. For each type of society, the level of the real wage rate will lie below the corresponding level of output per worker (average productivity of labor); that is, the line F, which shows the levels of labor productivity across societies, is accompanied by another line F' (not drawn) placed at a lower level of line F, which shows the corresponding levels of real wage rates. The gap in the levels of real wage rate across societies is related to the gap in the levels of average labor productivity. Given the relatively lower level of labor productivity in sigma society

compared to epsilon, there is no way for sigma society to have a relatively higher level of real wage rate compared to epsilon.

Therefore comparative cost advantage is determined by differences in the levels of labor productivity alone. Epsilon societies will specialize in the production of good B and sigma societies will specialize in good C. The differences in the level of real wage rate will be related to the differences in the labor productivity in the production of good B and good C in these two types of societies.

This trade model that is logically derived from the unified theory is similar to the well-known Ricardian theory of trade. The difference is that in our model, contrary to Ricardian models, labor productivity is not fixed exogenously and labor is not the only factor of production. However, what is fixed in this short run model of the unified theory is the *level* of labor productivity. Because the conclusions are similar to those derived from the standard Ricardian model, this trade model of the unified theory may be called the *generalized Ricardian model*.

The competing trade model is the one derived from neoclassical theory, which is called the Heckscher-Ohlin-Samuelson model. This model assumes that the source of comparative advantage is the difference in factor endowments among countries. This model predicts real wage equalization across countries, which is refuted by facts. The trade model of the unified theory does not predict real wage equalization, as in the Ricardian model. Moreover, empirical evidence shows that indeed countries trade goods according to the differences in labor productivity, as predicted by the Ricardian model (see the evidence in two of the most popular textbooks: Krugman & Obstfeld 2009, Figure 3-6, p. 49, and Carbaugh 2011, Figure 2-9, p. 57).

So far we have dealt with what is known as inter-industry trade. Another feature of the real world is intra-industry trade, in which the same goods move among countries (automobiles, textiles). This type of trade is explained in the standard trade literature by the existence of increasing returns to the firm or the industry (Krugman & Obstfeld 2009). However, this form of trade will be ignored here. Inter-industry trade is more significant in the relations between the First World and the Third World.

In the long run, what are the factors that determine international trade patterns? The answer will call for a theory explaining changes in labor productivity or output per worker across capitalist societies. According to the static model of the unified theory, labor productivity depends upon factor endowments, technology, and initial inequality. Thus the answer to the question lies in the process of capital accumulation and technological change, for a given level of initial inequality of countries. A dynamic model of the unified theory is then needed, which will be presented in the next chapter. The final answer will be provided in Chapter 13.

10.6 Additional Empirical Evidence

In order to falsify the empirical predictions of the static model of the unified theory with facts, measurements of the exogenous variables are needed. Regarding capital per worker differences, the available data indicate significant differences. Using the Penn World Table database, a study made the calculations of the value of capital stock per worker for a

sample of First World (8 countries) and Third World (7 countries) for 1997 and found that the ratio was 6:1; with South Korea the ratio was 2:1 (Carbaugh 2011, Table 3.2, p. 71). Another study, using different sources, and similar sample size for 1994, found similar ratio, 8:1 (Hofman 2000, Table 2, p.51).

On the other hand, workers in the First World are also endowed with higher levels of human capital, measured by years of schooling, than those in the Third World. The adult illiteracy rate was around 40% in the Third World compared to almost zero in the First World by 1998 (World Bank 2001, p. 277).

Regarding the inequality in the distribution of economic assets, the other exogenous factor, empirical studies are scarcer. A study on household *wealth inequality* presents estimates of Gini coefficients for a sample of 19 capitalist countries (16 from the First World and 3 for the Third World) for the year 2000, in which the average Gini coefficient for each group of countries is very similar, around 0.67 (Davies et al 2010, Table 7, p. 246). Another study published by Credit Suisse Research Institute (2011), showed that 85% of the world wealth is concentrated in the First World households.

On agricultural land concentration, one of the few studies about international comparisons, based on a sample of 103 countries of the world from the FAO database, for the period 1950-1990, showed estimates of Gini coefficients by regions. Considering only capitalist countries, the average Gini coefficients for the First World and the Third World were not much different, around 0.60 (Deininger and Squire 1998, Table 2, p. 266).

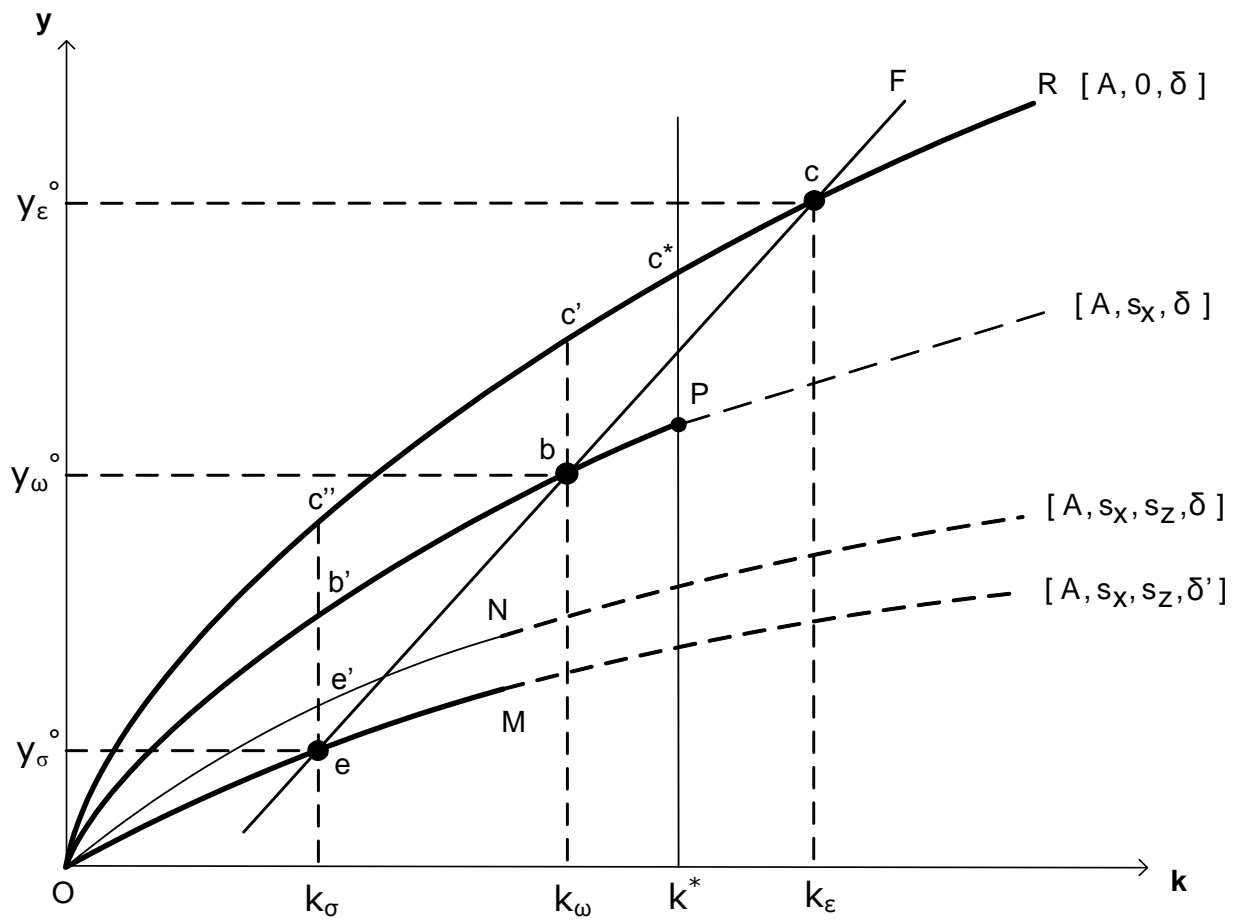
The concentration of human capital was estimated by this author from the international data of adult schooling years, for the year 1995, based on the UNESCO dataset, which is presented in a study by Barro and Lee (2000, Table 3, p. 12). The resulting average Gini coefficient from distributing the years of education between the adult population in the Third World countries is 0.60, whereas for the First World is only 0.28.

These estimates on inequality in the distribution of land, physical capital, and human capital indeed suggest that the overall inequality in economic assets is higher in the Third World compared to the First World, in which the major source of the inequality lies in the high concentration of human capital in the Third World. Differences in the concentration in the other economic assets do not seem to be significant. Human capital concentration is the key asset in the inequality of economic assets between the First World and the Third World.

Citizenship is a qualitative variable, which makes much harder to measure its degree of inequality. Formal rules (constitutions) establish equality of citizenship everywhere. The real question is the enforcement of those rules. Second class citizenship may in fact exist in the First World, but the model assumes that it applies mostly to Third World countries with strong colonial legacy. Empirically the concept of second class citizenship in the Third World applies to individuals whose ability to exercise rights is limited by informal norms. This is a colonial legacy because second class citizens are the descendants of the subordinated populations of the colonial history, which are called *z-populations* in the model. The international literature supplies some qualitative studies describing the different forms that second class citizenship take in the Third World and pointing out their importance for the entire economic and political processes (cf. Stewart 2001, 2008).

In sum, the empirical predictions of the static model are then consistent with the available empirical facts shown in Table 2.2, Chapter 2. The static model explains why income levels differ between the First World and the Third World. These countries operate at different levels of labor productivity curves, which is the result of the effects of two exogenous variables: factor endowments and initial inequality differences. Why do Third World countries are on average more unequal than First World countries? The static model also explains this fact: it is caused by the initial inequality differences. Therefore, the static model of the unified theory can explain the level component of Facts 6 and 7. The other component—persistence over time—needs a dynamic model, which will be presented in the next chapter.

Figure 10.1. Levels of Output per Worker by Types of Capitalist Societies



CHAPTER 11

GROWTH AND INEQUALITY IN THE CAPITALIST SYSTEM

In order to explain the persistence of differences in income levels between the First World and the Third World a dynamic model of the capitalist system is needed. Therefore, the process of economic growth now becomes the object of analysis. There is also the empirical fact that income distribution within countries is, on average, more unequal in the Third World countries than in the First World countries and that this inequality difference is also persistent, which need to be analyzed in a growth context.

The modern theory of economic growth is a well-developed field in standard economics. The growth theory to be presented in this chapter departs in several ways from the standard theory. But the most important difference is the search for a unified theory of capitalist development, which should explain both growth and distribution.

11.1 A Dynamic Model of the Unified Theory

In order to answer those questions a dynamic model of the unified theory is needed. The trajectories of both income level and income inequality of capitalist countries are the endogenous variables now and they need to be explained. While investment in physical and human capital was exogenously determined in the static models, now investment is endogenous and thus needs to be explained.

The construction of the dynamic model implies the introduction of auxiliary assumptions. The model assumes the existence of a global investment fund in the capitalist system, which will be allocated to individual societies. The global investment level is endogenously determined. In equilibrium, global investment will be equal to global savings. Domestic savings depend upon total output in each type of society; therefore, global savings will depend on global income. Given the initial output, savings will be determined, which will become the global investment fund, a flow variable. The global investment fund is then allocated to each type of society. Domestic investment in a particular society will represents a fixed proportion of total output; this proportion depends upon its relative degree of risk, which depends upon its relative degree of social disorder, which in turn depends on its relative degree of the initial inequality.

The equality between savings and investment need not hold true for each type of society, but only in the aggregate. The assumption is that investment in the sigma society is higher than domestic savings and hence it is financed with foreign savings. Sigma is a debtor society and no limits to foreign debt exist (foreign debt crisis is ignored). The supply of bank credit, which depends upon savings and the money supply, is also endogenous, and can be safely ignored. The saving-investment relation in the model is clearly a very simplified one.

The model thus assumes that the interactions between the different types of capitalist societies are reduced to the allocation of global investment. Foreign investment is assumed to have perfect mobility between societies. Contrary to the perfect mobility of capital, it is also assumed no mobility of workers. Hence, societies compete with each other in attracting investment, particularly investment in physical capital.

The endogenous variables of the dynamic model now include the growth rate of output per worker and the degree of income inequality in each type of capitalist society. The exogenous variables include the global rate of technological progress, and the rates of population growth, the investment/output ratios, the initial factor endowments, and the initial inequality degrees of the three capitalist societies.

Additional assumptions of the model include the following. It will be assumed that the dynamic equilibrium is a sequence of static equilibrium situations, which occurs at effective full employment in each society. Therefore, money is neutral and the relations between nominal and real variables may be ignored. The rate of excess labor supply remains constant over time, so that employment level grows at the same rate of population growth. Consequently, labor market will constitute the core of the system and its equilibrium is then all that is required as general equilibrium condition. For simplicity in the comparison of output per worker across societies, there will be only one single good produced in the capitalist system. (This is applying the Hicksian composite good principle: if the relative prices of a group of goods remain fixed, that group can be treated as a single good.)

On the production function, the model will assume constant returns to scale: double quantity of inputs produce double quantity of output. Therefore, if the quantity of machines and workers were double, total output would also double, which imply that output per worker remains unchanged. In this case, output per worker would not grow. Where does economic growth come from? If new technology were also included in the production process, then total output would increase by more than double, and output per worker would increase.

Technological progress means new knowledge in the input-output relations, which is the result of investment in research and development (R&D). The model assumes that the growth rate of technological progress is constant and exogenously determined; in addition, the model assumes that technological progress is labor-saving, which means labor productivity enhancing. The model also assumes a production function in which total output in society depends positively upon the quantity of capital (physical and human combined) and workers, in addition to technology. The growth rate of total output will thus depend positively upon the growth rates of each of these factors.

As to capital accumulation, the model assumes that investment in physical and human capital represents a fixed proportion of total output. So aggregating physical and human capital into a single composite good called simply "capital" implies measuring both in units of output. This annual investment is the addition to the stock of capital and thus this ratio (investment flow/stock) represents the growth rate of the stock of capital. Total output and investment will refer to net values, net of depreciation.

Economic growth will be defined as the continuous increase in output per worker over time. This concept of economic growth is intended to reflect increases in the labor productivity in society.

The growth process in each society operates as follows. In the initial period, society produces a certain amount of output per worker; investment in capital (physical and human) as a fixed proportion of total output also takes place. (Note that the investment ratio is not necessarily equal to the domestic saving ratio because, in an environment of perfect mobility of capital, investment is independent of domestic savings). Then in the second period, society starts with higher capital per worker, and new technological level, and thus produces a higher output per worker; again investment in capital per worker takes place but in a higher amount now. Then in the third period, society starts with a much higher capital per worker and a new technological level, and thus produces a much higher output per worker, and so on. Thus output per worker grows continuously over time.

Some basic numerical relations are introduced now. Growth rate refers to the percentage increase per unit of time, such as a year, from an initial value. Think of your savings deposit in a bank that is earning an interest rate per year. If the initial savings deposit was 100 dollars, and the interest rate is 10% per year, then at the end of the year the amount of the deposit is 110 dollars, at the end of the second year will be 121 dollars, at the end of the third year will be 133 dollars, and so on. The growth rate in this case is the interest rate. Note that the amounts at the end of the periods are not 110, 120, 130, and so on, which is due to the cumulative nature of growth.

11.2 A Dynamic Epsilon Model

The analysis starts with economic growth in the epsilon society. The economic structure is composed of a capitalist sector alone.

In order to derive more precise relations, the dynamic model of epsilon will assume the production function of the Cobb-Douglas type. In the capitalist sector, the only sector in the epsilon economy, the aggregate production function will take the following form:

$$Y = K^{\alpha} (AL)^{1-\alpha}, \quad 0 < \alpha < 1 \quad (11.1)$$

The term Y is total (net) output, K is the stock of capital, A is the level of technology, and L is the employment level. This production function incorporates the standard assumptions in growth theory, which are: (a) constant returns to scale and (b) technological change is labor augmenting. Given K , output depends upon the combination AL , where for given L , each worker produces A units; so an increase in the value of A is equivalent to the increase in the number of workers. Then A is the level of technology and its increase represents technological progress.

The standard procedure in the construction of a dynamic growth model is to construct a new variable $\tilde{L} = AL$, which measures labor in efficiency units or effective labor units (cf. Barro and Sala-i-Martin 2004). Then the production function shown in (11.1) can be transformed into another in which \tilde{L} , instead of L , is utilized. Hence

$$\tilde{k} \equiv K / \tilde{L} \quad (11.2)$$

$$\tilde{y} \equiv Y / \tilde{L} = \tilde{k}^{\alpha} \quad (11.3)$$

The steady state condition is that $\Delta \tilde{k} = 0$. If it were not, then the value of \tilde{k} would be changing until the steady state would be attained. This condition can then be written as

$$\Delta \tilde{k} / \tilde{k} = \Delta K/K - \Delta A/A - \Delta L/L = 0 \quad (11.4)$$

$$\Delta K/K = I/K = (S + S^*)/K = (1+a) S/K = (1+a) sY/K = eY/K, 0 < e < 1$$

$$e Y/K = \Delta A/A + \Delta L/L,$$

$$e \gamma / \tilde{k} = g + n$$

$$e \gamma = (g + n) \tilde{k}$$

The first equation in system (11.4) is the steady state or equilibrium condition. The second equation shows the determinants of capital accumulation in an open economy. It assumes that domestic investment (I) is a constant fraction of the global investment, which implies a fixed rate of investment to domestic total output; call this investment ratio e . Domestic investment will be equal to domestic saving (S) plus or minus foreign savings (S^*). Assume that $S=sY$ and, just for simplicity, that $S^*=aS$, then $e=(1+a)s$. The investment ratio e may be equal to the domestic savings ratio s (if $a=0$), but it is not necessary; in an environment of perfect international mobility of capital, domestic investment could be higher or lower than domestic savings. The third equation shows the condition that capital accumulation grows at a rate that is equal to the sum of the growth rates of labor and technology, as the steady state condition. The last two equations are just substitutions and algebraic manipulations.

The steady state condition is represented in Figure 11.1. The investment per effective worker is represented by the curve M; whereas the investment per effective worker required to keep the amount of capital per effective worker constant is represented by the curve N; the steady state condition implies that both curves must cross each other, which occurs at point \tilde{E} . The equilibrium values of \tilde{k} and γ are derived from the equilibrium condition. These values of the two endogenous variables will be repeated period after period forever, as long as the exogenous variables remain unchanged.

Transforming the steady state into dynamic equilibrium of output per worker, and using relations (11.2), (11.3), and (11.4), it follows that

$$\Delta k^*/k^* = \Delta A/A = g \quad (11.5)$$

$$\Delta y^*/y^* = (1 - \alpha) \Delta A/A + \alpha \Delta k^*/k^* = g \quad (11.6)$$

Capital per worker and output per worker will grow at the same rate: the rate of technological change, which is exogenously given. This is the requirement to satisfy the steady state. The implication is that the capital-output ratio (k/y) will remain constant in the growth process.

The dynamic equilibrium in units of output per worker and capital per worker is shown in Figure 11.2. Given the coefficients e , g , n , and the initial conditions of equilibrium A_0 and k_0 , such that $\tilde{k}^* = k_0/A_0$, the initial equilibrium takes place at point E^0 . Because capital has been accumulated, and new technology been adopted, in the next period the level of technology will be A' , which will shift the initial production frontier R to R' , which in turn will imply a shift of the investment curve from M to M' , and a new equilibrium will be reached at point E' . And so on. In this process the ratio y/k remains

constant. Thus, income per worker y^* grows endogenously along the points F^0 , F' , F'' , which constitute the dynamic equilibrium path.

The dynamic equilibrium can be written in functional form as follows:

$$\dot{y}^*(t) = f(t; e, n, g), f_1 > 0, f_2 > 0, f_3 < 0, f_4 > 0 \quad (11.7)$$

Given the initial conditions of equilibrium and the values of the exogenous variables, income per worker will follow a particular time path.

But the initial conditions \tilde{k}_0 may not be of equilibrium, such that $\tilde{k}_0 < \tilde{k}^*$. The transitional dynamics will take place to reach spontaneously the equilibrium value \tilde{k}^* . This is illustrated in Figure 11.3. The figure uses a proportional (logarithmic) scale, so that the each unit on the vertical axis corresponds to an equal percentage change in output per worker, and the slope measures a growth rate. The steady state equilibrium is given by the line M^*R^* . The level of the curve is determined by the exogenous variables (e, n, g); the slope in this graph indicates the rate of growth of output per worker, which is a constant, and equal to g , the rate of technological change. Let the initial output per worker condition be represented by point M_0 ; then the trajectory over time of y reaches the dynamic equilibrium trajectory, which occurs at period t_1 . In the transitional dynamics y grows at a higher rate than g .

It should be clear that there exists a relationship between the equilibrium conditions showed in Figures 11.1, 11.2 and 11.3. Given the equilibrium initial conditions, and given the exogenous variables, the steady state equilibrium, point \tilde{F} in Figure 11.1, maps into the segment F^0F'' in Figure 11.2, which in turn maps into line M^*R^* in Figure 11.3. In Figure 11.3, the line M^*R^* shows the dynamic equilibrium trajectory of income per worker. This line will be called here the *growth frontier curve*.

From any initial condition different from that of equilibrium, which will put the initial output per worker below the intercept of the growth frontier curve, the economy will move toward the frontier spontaneously, for the system is stable. This trajectory is called *transitional dynamics*.

Figure 11.3 can show the effects of changes in the exogenous variables upon the endogenous variable output per worker (y). An increase in variable e implies a new growth frontier curve, a shift from M^*R^* to $M^{**}R^{**}$; thus there will be a transitional dynamics, a move from a position out of equilibrium to the new equilibrium, which implies a growth rate of variable y that is temporarily higher than compared to the long run growth rate of the frontier curve. A decrease in variable n will have similar effect. An increase in variable g will increase the slope of the growth frontier curve. These effects are shown as the signs of the partial derivatives of equation (11.7).

It should be noted that the curve OM in Figure 11.1 represents the investment function, not the saving function. Changes in the coefficient e will thus shift the steady state position, but changes in s will not. If s increases, the OM curve will remain unchanged; but the consequence will be that foreign savings will decrease. The adjustment between domestic investment and domestic saving is made by foreign savings. The risks of

foreign debt crises are just ignored. In this open model, growth in the epsilon society is investment driven. (This is a difference with standard growth models in which growth is saving driven, for they assume a closed economy.)

Other predictions are derived from the transitional dynamics. Initial values of output per worker that are below of that of the growth frontier curve will generate a growth rate that is temporarily higher than that of the frontier; moreover, the lower the initial value, the higher the temporal growth rate.

It should be noticed that there are two growth rates of output per worker in the model. One is the long run, which refers to the frontier curve and is equal to the growth rate of technological change (g) and is exogenous; the other is the temporal growth rate, which refers to the transitional dynamics (g') and is endogenous, such that $g' > g$.

What happens with income distribution in the process of growth of epsilon society? Distribution refers to output shares going to capitalists and workers. Along the growth frontier curve, output per worker grows at the rate g . Output per worker is just the same as the average labor productivity. In the Cobb-Douglas production function, the marginal productivity of labor is a fixed proportion of the average labor productivity. Profit maximization of firms implies that real wages will be equal to marginal labor productivity; hence real wage will be a constant fraction of output per worker; that is, real wage will grow at the same rate of output per worker. Hence,

$$W/Y \equiv w/y = (\partial Y / \partial L) / y = (1 - \alpha) A^{1-\alpha} k^\alpha / A^{1-\alpha} k^\alpha = (1 - \alpha) \quad (11.8)$$

In the process of economic growth, the share of labor in total output will thus be constant and equal to $(1-\alpha)$; consequently, profits share will be equal to α coefficient.

Along the transitional dynamics, output per worker will grow faster than along the frontier curve; but marginal productivity will still be a constant fraction of the average productivity; hence real wages will grow at the same rate than output per worker. Income distribution will not change in this case either. In sum, income inequality in epsilon society will not change in the process of economic growth.

11.3 A Dynamic Omega Model

Omega is an overpopulated society. The marginal productivity of the total labor supply is equal to zero. The labor market is thus non-Walrasian. The economic structure of the omega model includes a capitalist sector together with a subsistence sector, in which the excess labor supply is able to generate income as self-employed. The general equilibrium condition between the two sectors is sequential, that is, the subsistence sector size is determined only once the equilibrium in the capitalist sector has been determined. Self-employment in the subsistence sector is residual, for it is equal to the excess labor supply. Another equilibrium condition is that the labor market operates with efficiency wages, which implies that the market wage rate must be higher than the opportunity cost of wage earners, which is given by the marginal income in the subsistence sector. Unemployment can safely be ignored.

Total output or national income in omega society may then be written as

$$\begin{aligned}
 Y &= Q + V & (11.9) \\
 Q &= K^\alpha (A D_h)^{1-\alpha}, \quad 0 < \alpha < 1 \\
 V &= K_{hs}^\beta L_s^{1-\beta}, \quad 0 < \beta < 1 \\
 L &= D_h + L_s
 \end{aligned}$$

Q is total output in the capitalist sector and, as before, the production function is represented by the Cobb-Douglas production function of constant returns to scale, in which D_h represents the quantity demanded of workers. V is total output in the subsistence sector, in which the production function is also subject to constant returns to scale, and L_s is the quantity of workers self-employed in the subsistence sector. Production in the subsistence sector is based on human capital alone; moreover, technological change depends upon the level of human capital. The other assumption is that physical capital is zero or constant. If physical capital is accumulated by the subsistence production unit, then it will become a capitalist firm. The total quantity supplied of workers is homogenous and equal to L , which is allocated to both sectors.

The behavior of the capitalist sector will be similar to the one presented in the epsilon society because the behavior of the capitalist sector is independent of the subsistence sector, which is the residual sector. Omega is therefore an epsilon economy together with a subsistence sector.

The capital accumulation process in physical capital takes place in the capitalist sector only. The long run equilibrium of the capitalist sector has now the problem that there exists excess labor supply. Therefore, in order to attain dynamic equilibrium, the accumulation of physical capital has to accomplish two tasks: to absorb the excess labor supply, which implies increasing capital per worker, and then to equip the growing population with the same capital per worker. When the dynamic equilibrium is reached, the excess labor supply will become equal to zero and the subsistence sector will have disappeared; thus omega society will become epsilon.

Omega society is just a “young” epsilon society, in which factor endowments are such that the initial capital per worker is below of that of the equilibrium value of a “mature” epsilon society. The only difference is that omega is overpopulation and thus transitional dynamics implies the coexistence of the capitalist sector with the subsistence sector. So there is a role for the subsistence sector in the growth process of capitalism: it is the residual sector where workers generate their incomes until overpopulation disappears.

The growth frontier of omega society is given by the growth frontier of the capitalist sector, which is the same as the frontier derived in the epsilon model. The transitional dynamics in the capitalist sector has the same property we found in the epsilon model. Figures 11.1, 11.2 and 11.3 shown above may then be seen as depicting the dynamic model of the capitalist sector of the omega society.

What is new in the omega dynamic model is the trajectory of the subsistence sector and with it the trajectory of the national income per worker. From the production function of the subsistence sector, third equation in the system (11.9), it follows that

$$v = k_h^\beta \quad (11.10)$$

$$\Delta v/v = \beta (\Delta K_h/K_h + \Delta L_s/L_s) \quad (11.11)$$

Output per worker in the subsistence sector v depends on the ratio of human capital per worker only. The assumption here is that technological change is generated by and for the capitalist sector. Therefore, the subsistence sector will have to make adaptations of the new technology before adopting it. The model assumes that these adaptations depend upon the level of human capital per worker, as Theodore Schultz (1975) suggested long time ago.

From equation (11.10) it follows that the growth rate of output per worker in the subsistence sector depends upon the difference between the growth rates of human capital minus that of self-employed workers; however, in omega society, the subsistence sector does not expand but it contracts due to the expansion of the capitalist sector. Then the growth rate of self-employed workers is negative, which results in the positive sign of self-employment in equation (11.11).

In sum, in the omega model, as the capitalist sector expands, the output per worker in the subsistence sector increases over time on two accounts:

1. Human capital accumulation takes place in society for all workers, who are homogenous labor, including for those workers engaged in the subsistence sector;
2. The number of workers diminishes in the subsistence sector and labor becomes more productive just due to diminishing returns (on reverse). The growth rate of output per worker in the subsistence sector is thus endogenous.

National income per worker will be equal to the weighted average of the output per worker in each sector. The levels between the two sectors are clearly different. The capital per worker ratio and the level of technology are both higher in the capitalist sector, which implies a higher level of output per worker in the capitalist sector compared to that of the subsistence sector.

In terms of growth rates, the capitalist sector grows at a rate that is higher than that of the growth frontier curve because it moves along its transitional dynamics. The subsistence sector will grow at a rate that is smaller than that of the capitalist sector. The comparison of equations (11.6) and (11.11) shows that the subsistence sector cannot grow at the same rate of that of the frontier curve, for it has to make adaptations of the new technologies; therefore, it will certainly grow at a rate that is smaller than that of the capitalist sector, which is moving along its transitional dynamics.

Figure 11.4 depicts the dynamic model of omega society. The vertical axis measures output per worker and the horizontal axis time. The growth frontier curve is given by the curve N^*R^* , which corresponds to the dynamic equilibrium of the capitalist sector. The segment NJ is the transitional dynamics of the capitalist sector, UV is the dynamic equilibrium of the subsistence sector, and HJ is the aggregate transitional dynamics of omega society. The subsistence sector disappears at period t^* . Beyond this period, omega society becomes endogenously epsilon society; that is to say, segment JR is the growth frontier curve of omega, now an epsilon society.

It should be noted that the average output per worker will grow at a rate that is higher than the growth rates of the two sectors. The reason is that the growth rate of the average value does not lie within the growth rates of the two sectors when the weights change over time, as is the case in this dynamic model. This is shown in Figure 11.4 by the figures inserted in the graph, in which curve NJ must grow at a faster rate than curve N^*R^* does because this is the only way to catch up with curve N^*R^* starting from an initial situation N, which is below point N^* ; similarly, curve HJ must grow at a faster rate than curve NJ does because this is the only way to meet at point J starting from point H, which is below point N.

On changes in income distribution, remember that omega society has three social groups that participate in national income distribution: capitalists, wage earners, and the self-employed. The initial income levels just follow that order. We already know that within the capitalist sector income distribution will not change; that is, the real wage grows at the same rate of output per worker in the capitalist sector. The only question is about changes between wage earners and the self-employed. The model predicts that output per worker in the capitalist sector will grow faster than output per worker in the subsistence sector, which implies that the self-employed (the poorest group) will become relatively poor, but also relatively smaller in size at the same time; hence, change in inequality will be ambiguous, which may safely be considered as inequality will remain unchanged.

11.4 A Dynamic Sigma Model

The economic structure of the sigma society includes a capitalist sector and two subsistence sectors. The model assumes as initial conditions two groups of workers: high human capital and low human capital. The latter workers are in addition second class citizens, with limited economic and political rights, such as limited access to public goods, and are called z- workers. The workers of the first group are also first class citizens, and are called b-workers (which was called x-workers in the static models and has been changed here for simple convenience of notation).

There exists labor market for workers endowed with high human capital only; so there exists a b-subsistence sector that is constituted by the excess labor supply of the first group. Z workers are out of the labor market because they are endowed with human capital that is too low to operate the modern technology in the capitalist sector and there is no incentive to incur in costly training, when at the same time there is much excess supply of skilled workers; thus they are self-employed in the z-subsistence sector. In sum, sigma society can be seen as an omega society together with a z-subsistence sector.

The equations representing total output in the three sectors of sigma society are

$$\begin{aligned}
 Y &= Q + V_b + V_z & (11.12) \\
 Q &= K^\alpha (A D_{hb})^{1-\alpha}, \quad 0 < \alpha < 1 \\
 V_b &= K_{hb}^\beta L_{sb}^{1-\beta}, \quad 0 < \beta < 1 \\
 V_z &= K_{hz}^\gamma L_{sz}^{1-\gamma}, \quad 0 < \gamma < 1
 \end{aligned}$$

$$L = (D_{hb} + L_{sb}) + L_{sz}$$

The sigma economy is thus composed of the omega sector and the z-subsistence sector. In this model, the z-subsistence sector is not connected to the rest of the economy via market exchange. Therefore, the omega sector will function separately from the z-subsistence sector; that is, sigma is a dual economy.

Conceptually, the analysis of the growth process in the z-subsistence sector is the only missing part to have the dynamic model of the sigma economy. As indicated by the fourth equation in the system (11.12), in the z-subsistence sector there is no physical capital or it is constant, for if physical capital were accumulated, the production unit would become a capitalist firm. Hence production relies mostly on human capital, which is accumulated through public investment, such as public education and health investments. This accumulation is thus financed through public funds and private savings of z workers. The model assumes that the accumulation of human capital is essentially determined by the quality and quantity of the supply of public goods, which is of second class quality, consistent with the political demands of second class citizens.

Economic growth in the z-subsistence sector implies technological modernization, substitution of traditional technology with modern technology. Technological change needs adaptation of the new technologies that were generated by and for the capitalist sector; moreover, the adaptation and adoption of new technologies depends upon the increase in human capital. Hence, from the fourth equation of the system (11.12), it follows

$$v_z = V_z/L_z = k_{hz}^\gamma \quad (11.13)$$

$$\Delta v_z/v_z = \gamma (\Delta k_{hz}/k_{hz}) = \gamma [\Delta K_{hz}/K_{hz} - n_z] \quad (11.14)$$

According to equation (11.3) output per worker in the z-subsistence sector depends upon the endowment of human capital per worker. The equation (11.14) shows that the growth rate of output per worker in the z-subsistence sector is equal to a fraction of the growth rate of human capital per worker, in which the growth rate of human capital (via public investment) is endogenous, but the growth rate of z-population is exogenous.

The model assumes that the accumulation of human capital in the z-population is not as rapid for the b-population, so no catch up will take place between the two types of workers. It also assumes that new technologies require even higher levels of human capital to operate them; hence, z-workers will remain in the z-subsistence sector because they are unable to reach the level of human capital of the b-workers. Comparing equations (11.6), (11.11), and (11.14), it follows that output per worker in the z-subsistence sector will grow at a rate that is the smallest among the sectors in the sigma society.

Figure 11.5 depicts economic growth in sigma society. Curve C^*R^* represents the growth frontier of the capitalist sector and curve CJ shows its transitional dynamics. Curves BB' and ZZ' represent the growth path of the b-subsistence sector and z-subsistence sector. Differences in the levels for the first two sectors were explained before in the omega model. The z-subsistence sector has the lowest level of output per worker because this sector has the lowest level of capital per worker and, consequently, the lowest technology

level (the most traditional technology) in sigma society. We may safely assume that the population growth rate of z-workers is equal to that of the b-workers. According to the sigma dynamic model, the growth rate of output per worker of curve C^*R^* is higher than that of curve BB' , which in turn is higher than that of curve ZZ' . Curve CJ grows faster than the curve C^*R^* .

The average output per worker of sigma society is given by curve SS' . The initial value at point S is just the weighted average of the values of output per worker in the three sectors, in which the value of the capitalist sector corresponds to that of the transitional dynamics. Up to period t^* , at which the b-subsistence sector disappears, the trajectory of the average output per worker is given by the segment SJ' . The curve C^*R^* cannot be reached because the z-subsistence sector continues and thus the average value must be below point J . The curve SJ' must grow faster than curve CJ because the gap between these two curves diminishes over time.

Beyond period t^* , the path is given by the segment $J'S'$, in which the growth rate switches to a lower level and then increases approaching the growth rate value of the curve C^*R^* . This is due to the effect of lower growth rate of the z-subsistence sector, which will tend to weaken over time. In the very long run, the average output per worker of sigma society will tend to grow at the same rate as does the capitalist growth frontier curve, but along a curve whose level lies below this frontier; hence, the two segments JR^* and $J'S'$ will tend to grow at the same rate. Consequently, in the process of economic growth, sigma society cannot grow along the capitalist growth frontier curve, even in the very long run. This is shown clearly in Figure 11.5.

In the process of economic growth, sigma society will eliminate the b-subsistence sector, which occurs at period t^* , but the z-subsistence will persist; therefore sigma society will be unable to become endogenously an epsilon society. In the very long run, the capitalists sector will coexist with the z-subsistence sector. The reason is that in the economic growth process, z-workers cannot become b-workers endogenously; z-workers are endowed with the lowest quantities of human capital and also with the lowest degree in political assets (as second class citizens). In the process of economic growth these initial conditions matter; particularly the accumulation of human capital can occur only through the supply of public goods. As second class citizens, z-workers have access to second class public goods only.

Given the unequal initial endowments, equalizing human capital between the z-population and b-population would imply a higher growth rate in the former, which the political system has no incentives in pursuing. In sigma society, human capital accumulation of social groups depends upon their initial economic and political asset endowments, which are unequal; therefore, the education system is not human capital equalizing, as shown in Chapter 9.

About distribution, two cases must be considered, before and after period t^* . In the first case, when the three sectors coexist, output per worker in the capitalist sector grows faster than that in the b-subsistence sector, which in turn grows faster than that in the z-subsistence sector. If the labor share of the sectors remained fixed, then income inequality would certainly increase; however, these shares change. The share of the b-subsistence sector decreases and that of the capitalist sector increases, which implies that the change in

inequality is ambiguous. (The initial and the new Lorenz curves may cross each other.) Hence, we may safely consider that inequality remains unchanged.

Beyond period t^* , output per worker in the capitalist sector grows at a higher rate than output per worker in the z-subsistence sector. Because the real wage rate will grow at the same rate of the former, the income gap between real wage and average income of z-workers will increase; moreover, inequality will increase because the poorest group will become relatively poor and its share in total population will be unchanged. In sum, the sigma model predicts that inequality will either increase or remain unchanged; but inequality decrease is not predicted in any event.

11.5 Dynamic Equilibrium of Growth and Distribution

The dynamic equilibrium of income per worker has been developed for epsilon, omega, and sigma economies, taken separately. Three partial theories have thus been presented through particular dynamic models. The question now is whether we have a unified theory that can explain both growth and distribution in the capitalist system, taken as a whole.

The result from the partial models is that there are only two growth frontiers, one corresponding to epsilon and the other to sigma. The growth trajectory of omega is just a curve showing the transitional dynamics to the frontier of epsilon. In this sense, omega and epsilon are qualitatively the same type of society, with quantitative differences only. The differences in factor endowments do not matter in the long run. In contrast, epsilon and sigma are quantitative and qualitatively different capitalist societies, for the differences in initial inequality matter in the process of economic growth.

The growth frontier curve of output per worker in each epsilon and sigma societies is given by the dynamic equilibrium of the corresponding capitalist sector. The frontier curve is determined by the same set of exogenous variables. The exogenous variables are the investment ratio (e) and population growth rate (n), which determine the level of the frontier curve, whereas the growth rate of technological change (g) determines the long run growth rate along the frontier.

Do these exogenous variables show differences between epsilon and sigma societies? If that were the case, different paths of growth and distribution would emerge in epsilon and sigma. In order to answer this question a unified model of the capitalist system, taken as a whole, is now presented.

The dynamic model of the unified theory will assume that global investment is endogenously determined. In equilibrium, global investment will be equal to global savings. Domestic savings will depend upon total output in each type of society; therefore, global savings will depend on global income, with a global saving rate (s^*). Given the initial global output, savings will be determined, which will become the global investment fund, a flow variable. The global investment fund will then be allocated to each type of society. The equality between savings and investment need not hold true for each type of society, but in the aggregate net foreign saving will be equal to zero; hence global investment will be equal to global savings as equilibrium condition.

These assumptions can be written as follows:

$$\begin{aligned} S &= s^* Y = I \\ s_1 Y_1 + s_2 Y_2 + s_3 Y_3 &= e_1 Y_1 + e_2 Y_2 + e_3 Y_3 \\ (s_1 - e_1) Y_1 + (s_2 - e_2) Y_2 + (s_3 - e_3) Y_3 &= 0 \end{aligned} \quad (11.15)$$

The first equation is the equilibrium condition between aggregate savings and aggregate investment. The second equation shows savings and investment in each society. The third just indicates that savings and investment need not be equal in each society, just in the aggregate. It follows that global savings is equal to the global savings ratio (s^*) multiplied by the global total income (Y), where the global savings ratio is just the aggregation of the savings ratios in each society. Exogenous changes in the global savings rate (s^*) will modify the global savings and thus the global investment, which will then be allocated into the different types of capitalist societies.

The unified model thus assumes that the interactions between the different types of capitalist societies are reduced to the allocation of the global investment. (The results will not change if the model assumes that the global society includes non-capitalist societies, as long as capitalist countries are competing for attracting global investment.) Foreign investment is assumed to have perfect mobility between societies. Contrary to the perfect mobility of capital, it is also assumed no mobility of workers. Hence, societies compete with each other in attracting investment, particularly investment in physical capital.

The basic endogenous variables of the dynamic model still include the growth rate of output per worker and the degree of income inequality in each type of society. The exogenous variables include the global saving rate (s^*) and the global rate of technological progress (g); among exogenous variables that are society-specific, the model includes the rate of population growth (n), the investment ratio e , the initial factor endowments (k_0), and the initial inequality degree (δ).

What are the factors that determine the investment ratio across different types of capitalist societies? The unified model will assume the following behavior of investors, which was developed in Chapter 8. Investors seek to maximize the mean rate of return and minimize risk of their investments portfolio. Factor endowments are such that capital per worker is higher in epsilon compared to sigma; but epsilon is also more endowed with human capital per worker, as professor Robert Lucas (1990) pointed out long ago. The first makes the mean rate of return lower (due to diminishing returns) in epsilon, while the second increases that return; hence factor endowments do not generate much difference in mean returns between these two types of societies. Thus the critical factor will be risk.

As to differences in risk, the model will assume that risk depends upon the degree of social order, which in turn depends inversely upon the degree of initial inequality, as shown in Chapter 8. The degree of initial inequality is lower in epsilon, which makes it a relatively lower risk society compared to sigma. Therefore, investors will have incentives to allocate much of their investment portfolio to epsilon society rather than to sigma; consequently, the investment ratio in physical investment will be higher in epsilon than in sigma.

Regarding investment in human capital, the unified model will assume that the investment ratio in human capital is higher in richer and more equal societies, in which

more public goods (in quantity and quality) to satisfy the demands for human capital of their more homogeneous citizens will be supplied. This is the case of investment in education and health. Therefore, the investment ratio in human capital will also be higher in epsilon than in sigma.

The investment ratio e has then been endogeneized. Hence, for society j

$$\begin{aligned} e_j &= I_j/Y_j = r_j I/Y_j = r_j (I/Y)/(Y_j/Y) = r_j s^*/m_j \\ &= \Phi(\delta_{0j}, s^*), \quad \Phi_1 < 0, \Phi_2 > 0 \end{aligned} \quad (11.16)$$

The first equation of the system (11.16) just follows the definitions and introduces new terms: r_j is the investment share of society j in global investment (I); Y is global output, and m_j is the share of society j in global output. This equation gives rise to the function Φ .

The unified model therefore predicts that the investment ratio (e_j) of capitalist societies depends inversely on the degree of their initial inequality δ_{0j} ; hence it is higher in epsilon societies, a more egalitarian society, than sigma. Then these differences in the investment ratio will imply that the level of the growth frontier curve of epsilon society will be placed in a higher position than that of sigma. The effect of the global saving ratio (s^*) is to shift the level of global saving and global investment for all capitalist societies, so the *differences* in the levels of the growth frontier are not affected by it.

As to the growth rate of population (n), the unified model will assume that this rate is lower in richer and more equalitarian societies. Consequently, the population growth rate is lower in epsilon than in sigma. This effect reinforces the difference between the levels of the growth frontier curves between epsilon and sigma that were established by the differences in the investment ratios.

In the unified model, therefore, the exogenous variables in the capitalist system as a whole have been reduced to only three: the global saving ratio (s^*), the growth rate of technological change (g), and the initial inequality of each society (δ). Because s^* and g are exogenous for all types of capitalist societies, differences in the levels of the growth frontier curve between epsilon and sigma can be explained by differences in investment ratios and population growth rates, which in turn depend upon their degrees of the initial inequality alone. More precisely, the gap between epsilon and sigma is due only to their differences in initial inequality. In explaining differences in income levels between epsilon and sigma, the reduced form of the system indicates that the initial inequality is the ultimate factor, whereas the exogenous variables e and n are just proximate factors.

The reduced-form equation of the dynamic model can then be written as follows. The equation for output per worker over time for society j is:

$$\begin{aligned} y_j^*(t) &= F^j(t; \delta_{0j}; s^*, g), \quad j = \epsilon, \sigma, \\ F_1 &> 0, F_2 < 0, F_3 > 0, F_4 > 0 \end{aligned} \quad (11.17)$$

This equation corresponds to the growth frontier curve, which in turn corresponds to the dynamic equilibrium of the capitalist sector in each type of society j , epsilon and sigma. This system of equations comprises the capitalist system, taken as a whole.

This result is shown in Figure 11.6, panel (a). The curve E^*F^* represents the growth frontier in the epsilon society, whereas S^*R^* corresponds to that in sigma society. The growth rate along both frontiers is equal to g , the rate of technological change.

Consider now the transitional dynamics. For each society, it can be written as

$$y^\epsilon(t) = f^\epsilon(t; k_0), f_1 > 0, f_2 > 0, \text{ where } k_0 < k_0^*(\epsilon) \quad (11.18)$$

$$y^\sigma(t) = f^\sigma(t; k_0, \xi_0), f_1 > 0, f_2 < 0, f_3 > 0, \text{ where } 0 < \xi_0 < 1 \quad (11.19)$$

Equations (11.18) and (11.19) refer to transitional dynamics in epsilon and sigma societies, which are separate for each type of society, for they have different frontiers. These equations ignore the initial level of technology (A_0) by assuming that it is associated to the initial level of capital per worker (k_0), as new technologies will somehow be incorporated in the new capital goods in the capital accumulation process. In the epsilon society, therefore, the initial capital per worker (k_0)—when it is smaller than that of the initial equilibrium condition ($k_0^*(\epsilon)$)—is the determinant of the transitional dynamics. In the sigma society, the transitional dynamics does not refer to the capitalist sector, but to the total output per worker or national income per worker. Its determinants include the initial capital per worker ratio (k_0) and the initial share of z -population in total population (ξ_0 , Greek letter zeta).

The transitional dynamics are also shown in Figure 11.6, panel (a). Any epsilon society with initial conditions of capital per worker and technology level lower than the equilibrium initial values will move towards the frontier curve E^*F^* , which implies a *growth rate* that is faster than that of the frontier (g). The segment AB' shows transitional dynamics. Any omega society will also move towards the frontier curve E^*F^* . This is represented by the segment CB , which is also transitional dynamics. At period t_1 , omega has eliminated overpopulation and has become endogenously an epsilon society.

Sigma society will also tend to move from an initial condition that is not of equilibrium towards the growth frontier curve S^*R^* , as indicated by the segment MM' , which shows transitional dynamics, which implies a *growth rate* that is higher than that of the frontier (g). This path refers to national income per worker and implies that the capitalist sector of sigma moves along its own transitional dynamics and reaches its frontier S^*R^* at period t_2 , when the b -subsistence sector has been eliminated, but the z -subsistence sector continues. Beyond period t_2 , national income per worker is equal to the weighted average of the output per worker of the capitalist sector (moving along the frontier S^*R^*) and that of the z -subsistence sector (which moves along a path that is located at a lower level, not shown) and its trajectory is given by the segment $M'M''$. The growth rate along $M'M''$ is smaller than that along MM' . As shown in the graph, the frontier curve will never be reached.

Changes in income distribution in the process of economic growth have already been determined above for epsilon and sigma. They can be summarized as follows. In the

epsilon society, income inequality remains unchanged in the process of economic growth, that is, along the frontier curve and along the transitional dynamics. (In the omega society, in transition to epsilon society, inequality will also tend to remain unchanged.) In sigma society, before period t_2 , inequality will also tend to remain unchanged; beyond this period, inequality will tend to increase, but this is not an empirically relevant case (never observed). Hence, income inequality tends to remain unchanged in sigma society as well.

On distribution, the reduce-form equations for the degree of inequality (D) can then be written as follows:

$$D^{\varepsilon*}(t) = G(\delta_0), G' < 0 \quad (11.20)$$

$$D^{\varepsilon}(t) = D^{\varepsilon*}(t) = G(\delta_0), G' < 0 \quad (11.21)$$

$$\begin{aligned} D^{\sigma}(t) &= H(\delta_0), H' > 0, \text{ where } t \leq t_2, \\ &= h(t; \delta_0, k_0, \xi_0), h_i > 0, \text{ where } t > t_2 \end{aligned} \quad (11.22)$$

For epsilon society, equation (11.20) shows the degree of income inequality along the growth frontier curve: the inequality of income flows depends upon the inequality in the distribution of the initial assets alone; then equation (11.21) indicates that the degree of income inequality along the transitional dynamics is the same as that along the frontier.

For the case of sigma society, equation (11.22) presents the degree of income inequality along the transitional dynamics (remember that in sigma the transitional dynamics curve does not reach the frontier curve). In sigma, therefore, income inequality outcome is more involved. There are two stages. In the first stage, when the sigma economy operates with the capitalist sector and the two subsistence sectors, up to period t_2 , the initial inequality determines the level of income inequality, which tends to remain constant over time. In the second stage, beyond period t_2 , when only the capitalist sector and the z-subsistence sector coexist, the *level* of the degree of income inequality also depends positively on the initial inequality, but also positively on the initial factor endowment and on the initial share of z-population in total population; regarding the *slope*, inequality tends to increase over time. But this second stage of sigma society has never been observed and can be neglected when submitting the model to empirical consistency.

Figure 11.6, panel (b), depicts the predictions of the unified model. It includes differences in the level of inequality by types of capitalist societies and also changes in the degree of inequality in the process of economic growth.

The empirical predictions of the unified model can be summarized as follows: First, the long run income level gap between epsilon and sigma societies depends upon the difference in their initial inequalities. The lack of output per worker equalization will persist over time as long as the initial inequality difference remains unchanged. However, income levels equalization will tend to occur between omega and epsilon. Second, the long run degree of income inequality in each type of capitalist society (epsilon, omega, and sigma) depends on its initial inequality, which does not tend to change in the process of economic growth. Hence, the order of income inequality in each society follows the order of the initial inequality; that is, the degree of income inequality is the highest in sigma and the lowest in epsilon, while omega lies in between; consequently, this order in inequality tends to persist in the process of economic growth.

Taking the capitalist system as a whole, the *global* degree of income inequality in the capitalist system will be the result of within-society inequality and between-society inequality. The within-society inequality depends upon each society's initial inequality, as shown above. The between-society inequality, which is equal to the gaps between the levels of the growth frontier curves, also depends upon the initial inequalities. In sum, *the global degree of income inequality in the capitalist system depends upon the global initial inequality in the distribution of economic and political assets*. Therefore, exogenous increases in the growth rate of technological change, which increases the long run growth rate of output per worker in each society, will not reduce the global income inequality, but exogenous reductions in the global inequality in assets distribution will create a more egalitarian capitalist system.

11.6 Empirical Evidence

The predictions of the dynamic model of the unified theory refer to between-country inequality and within-country inequality. They are represented in Figure 11.6. These predictions are consistent with Facts 6 and 7, listed among the basic empirical regularities of production and distribution in the capitalist system, in Chapter 2.

In Table 2.2, Chapter 2, output per worker is measured by GNI per capita (also called here "income level") and the degree of income inequality by the Gini index. Income level differences between the First World (epsilon societies) and the Third World with strong colonial legacy (sigma societies) are large and have increased even more (from 4.0 to 6.4 times) in the three decades for which comparable information is available; on the other hand, the income level differences between the First World and the Third World with weak colonial legacy (omega societies) are not as large, and the gaps has declined from 2.5 to 2.0 times. These facts are consistent with the first prediction of the unified model.

Regarding within-country inequality, Table 2.2 shows that the average degree of income inequality between 1950-2008 is the lowest in the First World countries and is the highest in the Third World countries with strong colonial legacy, with the Third World with weak colonial legacy lies in between; on the other hand, this order has remained unchanged in the last four to five decades. The stability of the degree of inequality appears to hold true if we compare the average of the period 1950-1970 with those of the entire period. These facts are also consistent with the second prediction of the model.

Further empirical consistency of the model predictions is now presented. On the persistence of income level gaps, the classic empirical study on convergence by Barro and Sala-i-Martin (2004) found that one pattern in the cross-country data is that the growth rate of real per capita GDP from 1960 to 2000 is positively correlated (although slightly) with the level of per capita GDP in 1960 in a sample of near 100 countries. If there were convergence towards a unique growth frontier, then there would be strong negative correlation: the poorer countries should grow at higher rates, which is not the case. The other pattern is that convergence does exist within the OECD countries, which largely corresponds to our empirical definition of First World (epsilon societies); hence, according to the unified model, we could say that First World countries tend to converge to the

common growth frontier, as the unified model predicts. But Third World countries have different growth frontiers.

The unified model predicts that omega societies will catch up with epsilon societies. Do Third World countries with weak or no colonial legacy tend to converge to the income level of the First World countries? Table 2.2, Chapter 2, answered affirmatively this question. According to economic historian Angus Maddison (1995), only Japan is the historical case of catching up in the long run. A country that in 1820 was among the group of poor countries, Japan has become a full member of the club of the First World countries. The other possible cases for the near future include only South Korea and Taiwan. In light of the unified model, these three Asian countries indeed started capitalist development as omega societies; therefore, these countries' growth performance cannot be seen as "miracles," as they are usually seen, for the unified dynamic model predicts this behavior.

In sum, the unified theory of the capitalist system predicts that First World countries and Third World countries with strong colonial legacy grow along different paths. They have different growth frontiers. In the process of economic growth, these two groups of countries will converge to their respective growth frontiers. In the growth process, therefore, the Third World will not catch up with the First World endogenously, as long as the exogenous variable—the inequality in the distribution of economic and political assets—remains unchanged.

The neoclassical growth theory also predicts no absolute convergence, only conditional convergence. Countries with a similar steady state will converge to this steady state, and the speed of this convergence relates inversely to the distance from the steady state (Barro and Sala-i-Martin, 2004). The prediction of the unified theory is similar. However, the exogenous variables are different. The savings rate and the population rate are exogenous in the neoclassical model, but endogenous in the unified theory. The ultimate factor that determines the steady state (or growth frontier curve) is the initial inequality in the unified theory.

A more recent study on economic growth by Oded Galor (2011) analyzes the growth process of the world economy, from the dawn of civilization to today, in a unified growth theory. It is "unified" in the sense of seeking to explain growth as one single process, one single history, not by epochs. It argues that inequality between countries (capitalists and non-capitalists) is explained by the differential timing of the take-off from stagnation to growth, which triggered a divergence in income levels across countries. The differential timing in turn is explained by country-specific factors (including here inequality and factor endowments, among *many* other factors). The institutional legacy of European colonialism as a deleterious factor is also recognized. The implication of this theory is that the current absolute divergence is only temporary.

On the prediction of stability in income inequality within countries, one thing needs to be clarified. We observe that most countries of the Third World operate with the coexistence of the three sectors of the sigma model, which implies that what we are observing is basically the transitional dynamics in these countries. For this case ($t < t_2$), the sigma model predicts stable or viscous income inequality over time in the Third World. The epsilon and omega models also predict this behavior for the First World countries and the Third World countries with weak colonial legacy. Thus the unified model predicts stability in income inequality across capitalist countries.

The classic econometric study by Li, Squire and Zou (1998) presents results that are consistent with such prediction. Using the income inequality database of the World Bank, and based on a sample of 49 countries, covering the period 1947-1994, they found that only seven countries showed statistically significant and quantitatively important time trend, that is, 42 countries showed stable Gini coefficients over time. The authors point out that the sample includes many countries with few observations, as small as four, which may make the statistical testing inaccurate (p.33). Replicating the regression analysis with a sample of 21 countries with 10 or more observations, they found similar results: out of 21 countries in the sample, 17 countries showed stable Gini coefficients over time.

Using a more recent database on income inequality constructed at World Bank by Branko Milanovic (2010), which covers the period 1950-2008, this author was able to update the results of Li et al. Considering only capitalist countries, and only those with at least nine observations, and using a homogeneous definition of income (“net income”) from the household surveys, the resulting sample size is equal to 24 capitalist countries (16 from the First World, two from the Third World with weak colonial legacy—South Korea and Taiwan—and six from the Third World with strong colonial legacy). The results of the regression model that replicates the same regression model of Li et al. are presented in Appendix B. Out of the 24 countries in the sample, 21 countries (88% of the sample) showed stable Gini coefficients over time. The three countries with significant time trends include Finland (negative trend), France (negative), and New Zealand (positive).

A simple comparison can be made between these results and those of Li et al. In their regression analysis based on 49 countries with four or more observations, 43 were capitalist countries. If from this new subsample of capitalist countries, we take only those countries with 10 or more observations, the sample size is equal to 16 capitalist countries, of which 14 (88% of 16) showed stable Gini coefficients over time. Thus one could say that the two studies show similar results.

Using a database other than the World Bank’s, Atkinson (1996) found similar results. In a sample of 17 countries of the OECD for the period 1970-1992, he also found no significant changes in income inequality over time, except for the United Kingdom.

At this point we need to recognize that in the Third World income inequality data are scarce, quantitatively and qualitatively. The quantitative limitations are clearly shown above in the database of household surveys: the sample size of countries is reduced drastically when the number of observations per country is increased. Qualitatively, it is known that profits and top salaries are not well captured in the Third World household surveys. (This is another characteristic of a sigma society, in which economic elites simply refuse to participate in surveys.) Income inequality from these surveys seems to measure mostly the distribution of labor incomes rather than national income (which includes profits).

A study on factorial income distribution (calculated through the method of national accounts of United Nations) utilized a sample of 12 countries from the Third World and also 12 countries from the First World and found that the average profit share in national income was 22% in the former and 24% in the latter, around 1990 (Gollin 2002, Table 2, p. 470). In a more recent study, Atkinson et al (2011) estimated the share of the top 1% from income tax and national accounts data for a sample of 14 countries (11 from the First

World and three from the Third World). The shares ranged from 8% to 17% in 2005. As to trends for the period 1950-2005, the results of this study were diverse: the share increased in four countries, declined in five, and remain more or less constant in two.

To be sure, in the data shown in Table 2.2, profits are mostly included in the Gini index for the First World countries, but they are not for the Third World; hence, income inequality differences are underestimated. To make them comparable, the Gini index for the Third World would have to be recalculated attributing income from profits, say 22% of national income to, say, the 0.1% of households!

The consequences of these limitations regarding inequality data in the Third World for the predictions of the unified model would be two fold. As to differences in the level of inequality, the degree of income inequality in the Third World countries would tend to be underestimated; therefore, the real gap between the Gini index for the First World and the Third World shown in Table 2.2 above is underestimated. As to stability on inequality trends over time, the estimated empirical changes in national income inequality in the Third World are to some extent uncertain because the database refers mostly to changes in the labor income part only.

11.7 Conclusions

This chapter has presented a unified theory of capitalist development, which intends to explain the determinants of both growth and distribution in the capitalist system. The ultimate factor that explains differences in growth and distribution paths between the First World and the Third World countries is their differences in the initial inequality in the distribution of economic and political assets.

The persistence of income level gaps between the First World and the Third World and the existence and persistence of gaps in the average degree of within-country inequality between the First World and the Third World (Facts 6 and 7 indicated in Chapter 2) are all explained by the models of the unified theory. Therefore, given the initial inequality in the distribution of economic and political assets, the First World and the Third World have followed different growth paths and inequality paths. These paths have implied persistence in within-country and between-country inequalities in the capitalist system.

According to the unified theory, therefore, there is path dependence in the process of capitalist economic development; that is, history matters. The initial inequality in each society does not become reduced endogenously in the process of economic growth; that is, differences in the initial inequality between societies are not endogenously eliminated. Initial conditions matter. The long run trajectory of the capitalist system depends critically on the history of the system.

In sum, facts do not refute the predictions of the models of the unified theory. Hence, there is no reason to reject these models and then we may provisionally accept the unified theory of growth and distribution at the present stage of our research until new empirical data or superior economic theories appear.

The objective of a unified theory still confronts a new challenge. It must explain the role of the biophysical environment in the growth and distribution process. The environmental problem is one of the fundamental problems of our time. The unified model presented here assumes that growth can go on forever, as it ignores the role of the environment. Explaining the relationships between growth, distribution, and the environment should be pursued by any unified theory of modern economics. This is the objective of the next chapter.

Figure 11.1. Steady State Equilibrium in Epsilon Society

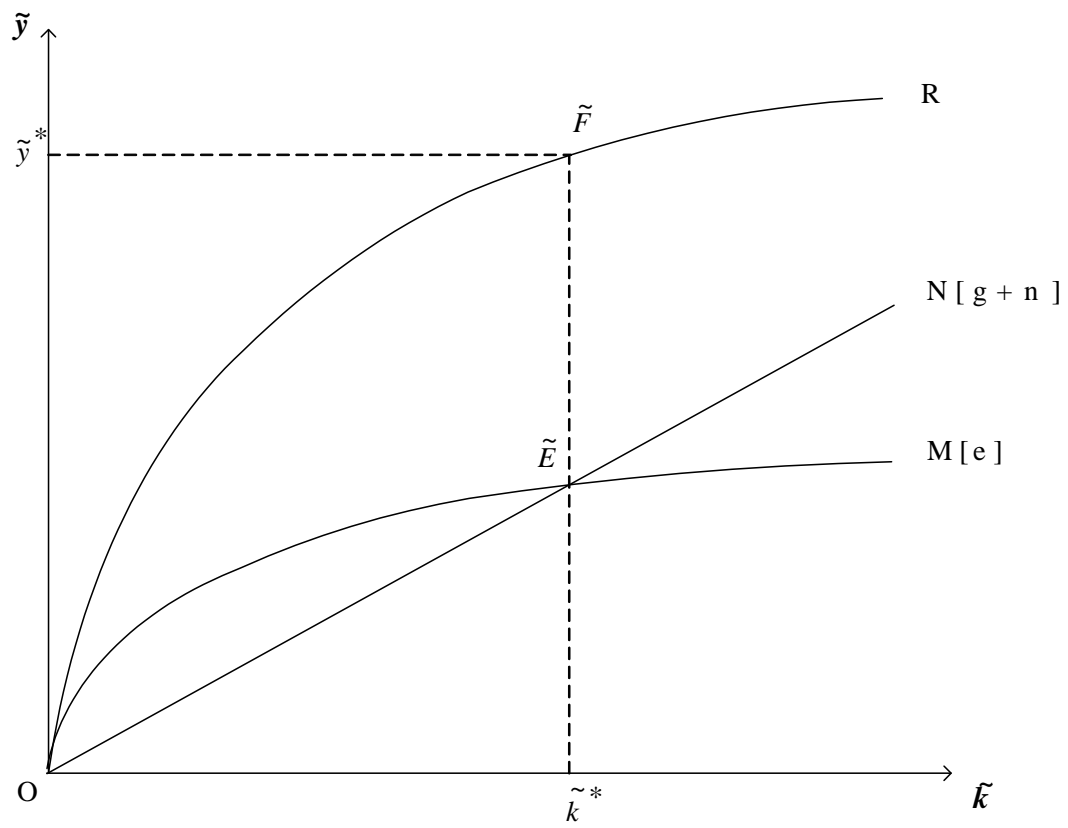


Figure 11.2. Output per Worker Dynamic Equilibrium in Epsilon Society

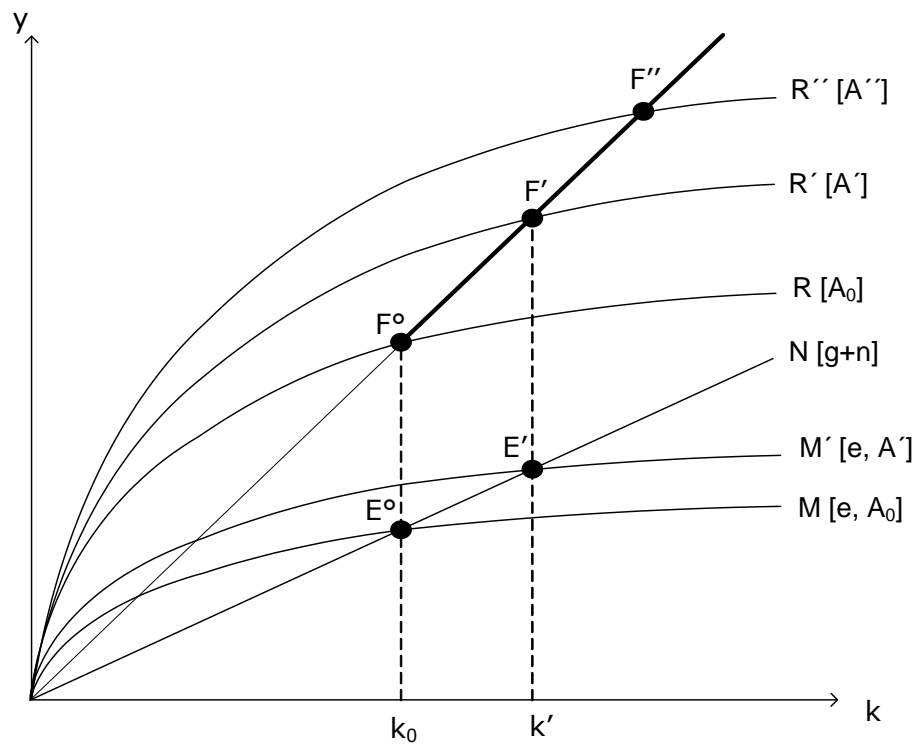


Figure 11.3. Output per Worker Dynamic Equilibrium and Transitional Dynamics in Epsilon Society

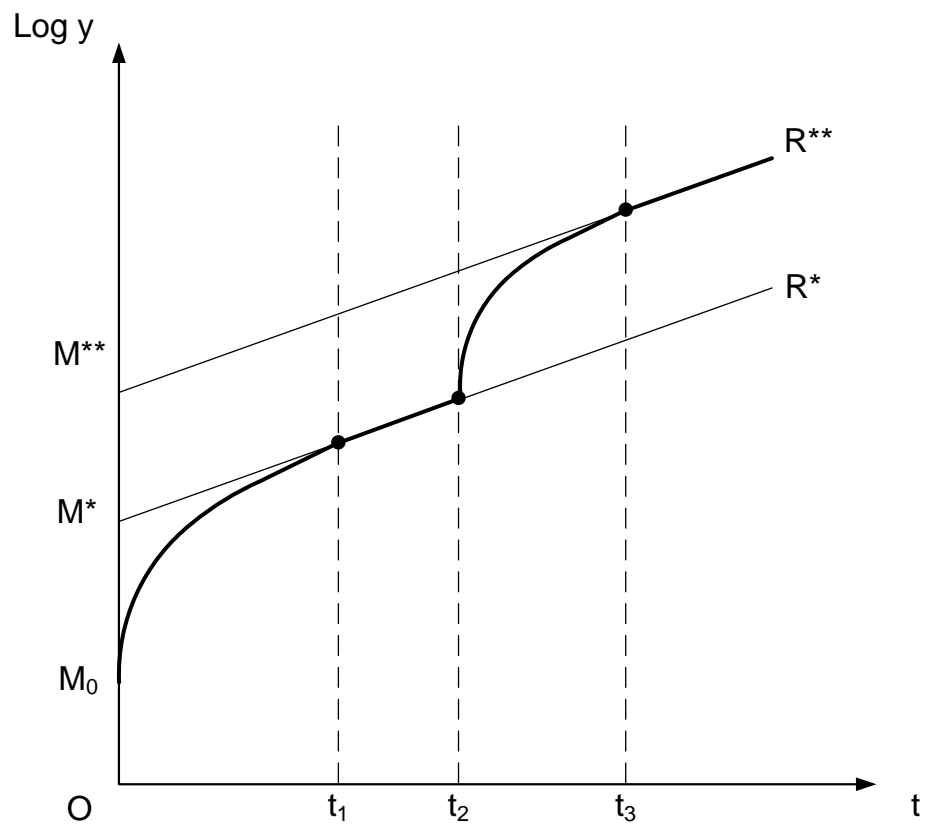
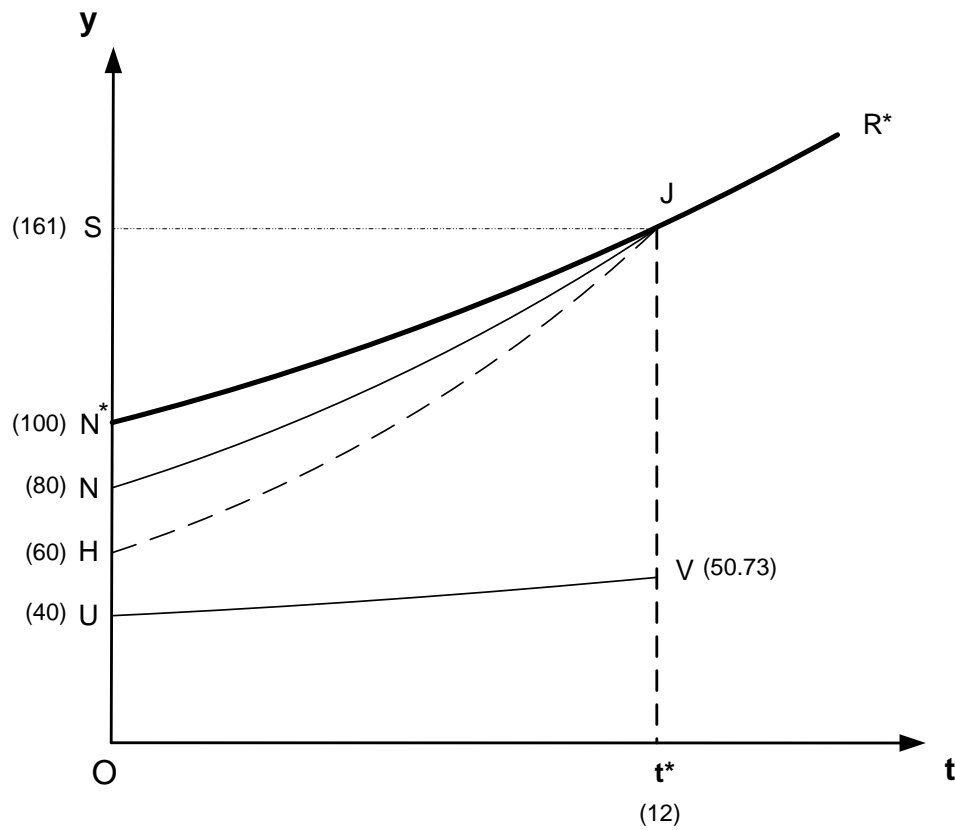
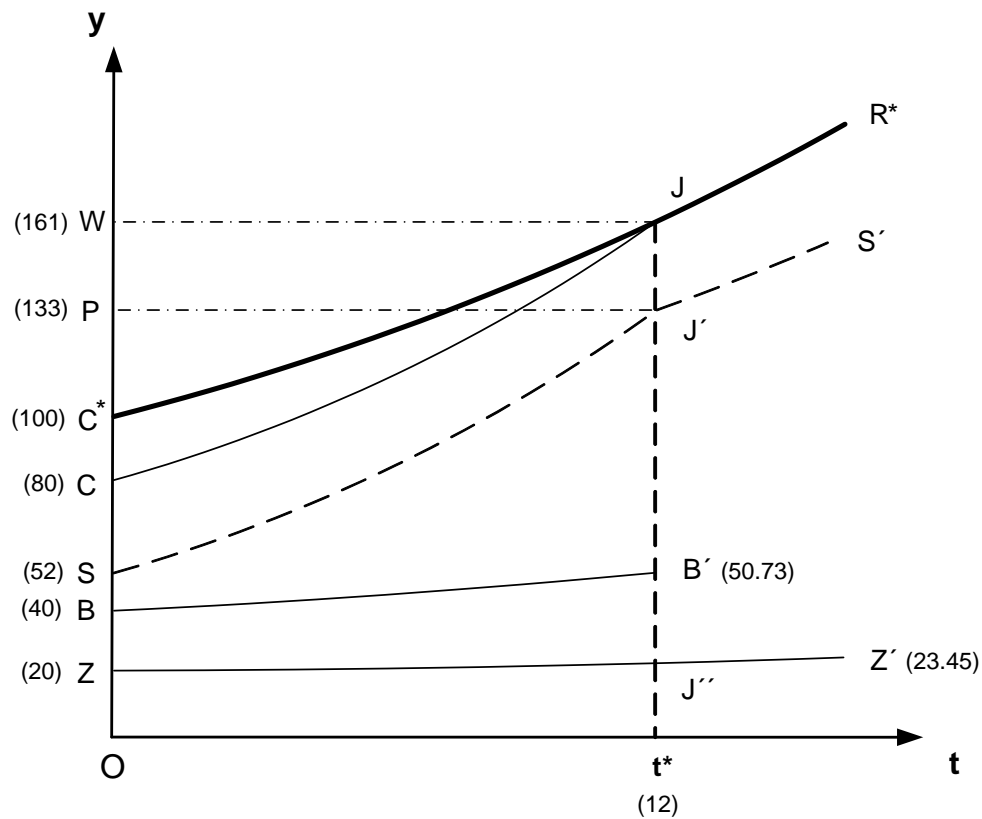


Figure 11.4. Growth in Omega Society



Note. The curves are drawn on natural scale. Initial conditions are indicated in the vertical axis, which measure initial values of output per worker (y). The initial labor shares are: 0.5 (capitalist sector) and 0.5 (subsistence sector). The curves assume constant annual growth rates, which are: $N^*R^*=4\%$ (capitalist growth frontier), $NJ=6\%$ (capitalist transitional dynamics), and $UV=2\%$ (subsistence sector). Curve HJ is the derived aggregate output per worker, with initial value $H=60$ and average growth rate equal to 8.6% . The subsistence sector disappears at period 12.

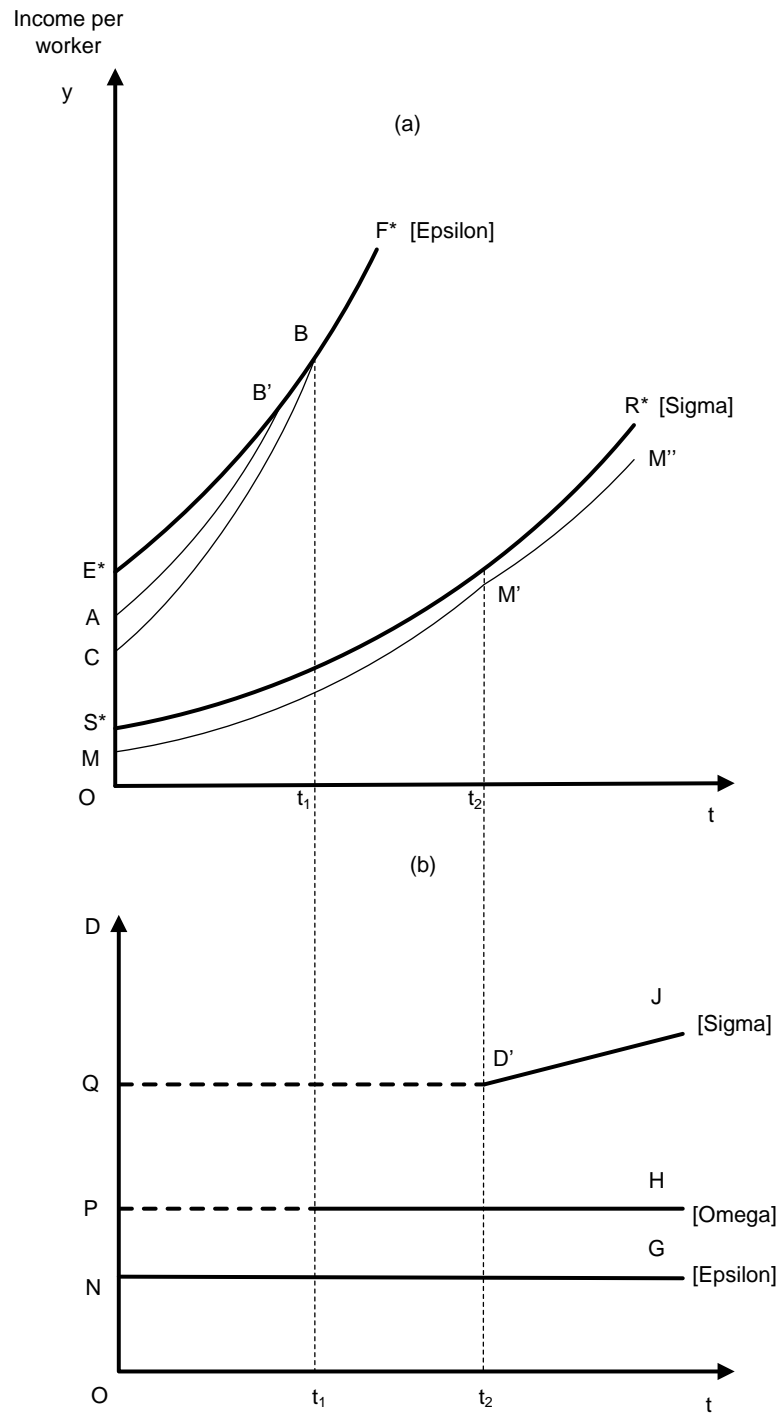
Figure 11.5. Growth in Sigma Society



Note. Curves are drawn on natural scale. Initial conditions are indicated in the vertical axis, which measure initial values of output per worker (y). The initial labor shares are: 0.4 (capitalist sector), 0.4 (subsistence sector X), and 0.2 (subsistence sector Z). The curves assume constant annual growth rates, which are: $C^*R^*=4\%$ (capitalist growth frontier), $CJ=6\%$ (capitalist transitional dynamics), $BB'=2\%$ (b-subsistence sector), and $ZZ'=1\%$ (z-subsistence sector). Curve SS' is the derived aggregate output per worker, with initial value $S=52$ and average growth rate along SJ' equal to 8.2% and that along $J'S'$ approaching asymptotically to 4% (the *growth rate* of curve C^*R^*). Subsistence sector X disappears at period 12.

Figure 11.6. Growth and Distribution in the Capitalist System

Note. Growth frontier curves E^*F^* and S^*R^* in Panel (a) are drawn on scale. Values in the vertical axis are: $E^*=30$ and $S^*=8$. Annual growth rates are: $E^*F^*=3\%$ and $S^*R^*=3\%$.



CHAPTER 12

GROWTH AND INEQUALITY UNDER ENVIRONMENTAL DISTRESS

In the dynamic model of the unified theory presented in Chapter 11, natural resources were ignored in the economic process. The implicit assumption was that natural resources were abundant. One of the regularities of capitalism (Fact 8), however, says that the biophysical environment has been degraded in the process of economic growth. In order to explain this fact, we need to introduce natural resources into the economic process.

Static and dynamic models have been presented so far. Static equilibrium implies that, given the values of the exogenous variables, the values of the endogenous variables can be repeated period after period *forever*. Dynamic equilibrium in turn implies that, given the values of the exogenous variables, the values of the endogenous variables will move over time along a given trajectory *forever*. In both cases, the economic process is seen as a mechanical process, just as the pendulum in physics.

In the previous chapter, the process of economic growth was seen as a mechanical process, for economic growth could proceed forever. There were no limits to economic growth. In this chapter, the dynamic economic process will be seen as an evolutionary process, in which not only quantitative changes take place in the process of economic growth, but also qualitative changes appear in that process. In the construction of an evolutionary theory of economic growth, the chapter will follow the approach initiated by the late Professor Nicholas Georgescu-Roegen (1971), in which the laws of thermodynamics (dealing with matter and energy) are included in the economic process.

The chapter will present a model of the unified theory in which natural resources play a significant role in the process of economic growth. Several intermediate models will be constructed to understand this complex system step by step.

12.1 Towards an Evolutionary Model of the Unified Theory

In order to apply process analysis, represented in Figure 1.1, Chapter 1, to the case of the economic process with natural resources, some new concepts are now introduced. Regarding input-output elements, there are two types of elements crossing the boundaries of the process: those that go into the process *and* come out of the process, and those that either enter *or* come out. The first group is called *fund factors* and the second *flow factors*. Fund factors include machines and workers, while flow factors include the material inputs and the material outputs.

In the production process, therefore, material input flows are transformed into material output flows with the help of fund factors, the agents of change. The underlying mechanism of this transformation is technological knowledge and social institutions.

The economic process is not an isolated, self-sustaining process when natural resources are taken into account. The economic process now interacts with the environment. Natural resources provide the flow factors, which come from the environment and cross the boundary of the economic process; then waste crosses the boundary from inside the economic process and goes into the environment, and so on. The idea is that the environment and the economic process constitute an integrated system. A model is now needed to establish those relationships more rigorously. In fact, several models will be needed.

The models will assume that the production process has two types of components: funds and flows. Funds and flows constitute two different categories and play different roles in the production process. Fund factors include machines and workers. Machines are made of good B, the only good produced in society. (Think of a society living with cereals only, which satisfies all human necessities, including cakes and liquor!) Flow factors include the proper output, which refers to the good produced, good B (a material good); the unintended output, which refers to the waste; and raw material inputs, which come from natural resources. Two categories of natural resources will be distinguished: renewable (biological) and non-renewable (called here mineral resources).

The economic growth process will be seen in this chapter as an evolutionary process, in which *Time* T is historical time (with past, present, and future). This is contrary to the view of economic growth as dynamic process (shown in the previous chapter) in which *time* t refers to mechanical time: economic growth moves in the same way irrespective of when the event occurs in historical time (just like a pendulum movement, which is invariable with respect to historical time).

The models will also assume a single world human society, which is endowed with stocks of machines and workers; it is also endowed with stocks of natural resources, distributed across the planet Earth. Just to oversimplify, the world society will be equal to world capitalism. Three models will be presented step by step, so as to construct a very comprehensive final model.

12.2 Model A: Economic Process with Non-renewable Natural Resources

The theoretical model will assume an abstract human society with asset endowments unequally distributed among individuals. Assets will refer to economic resources (physical and human capital) and political entitlements (degree of citizenship). It also assumes that, in the long run, the inequality in the distribution of the output flow is determined by the initial inequality in the distribution of economic and political assets, which is the result obtained in Chapter 7 above.

The algorithm applied in the construction of the proper model starts with model A, which assumes a particular production process in which good B is the only good produced. It is represented in the form of a *production system*, as follows:

$$Y^*(T) = F(K(T), L(T)) \quad (12.1)$$

$$Y^*(T) = (1/z) N(T), \quad z > 0 \text{ and } \sum N_j \leq S_0, \quad j=1, 2, \dots, T \quad (12.2)$$

The production system (12.1)-(12.2) assumes that the flow of gross output Y^* is produced in period T with the use of quantities of two types of production factors: the fund of services contained in the stocks of capital K and labor L —equation (12.1)—and the flow of material inputs N coming from the stock of non-renewable natural resources S_0 , which here will refer to minerals in the Earth's crust, taken as a single material—equation (12.2). Capital stock K is made of good B . Total workers who participate in the production process as wage earner and self-employed is denoted by L ; hence unemployment is just ignored. Renewable natural resources will be ignored for the time being.

The production system (12.1)-(12.2) represents a particular notion of production process in which $Y^*(T)$ is the output flow produced in period T and $N(T)$ is the flow of mineral resources used in the same period so that the output flow can be *repeated period after period* as long as the stocks of K and L remain fixed over time, and as long as the flow of mineral inputs are still available, given the initial stock S_0 .

The production system (12.1)-(12.2) assumes *limitational technology*; that is, the first type and the second type of factors are not substitutable to each other. Mineral resources cannot be substituted by capital or labor; however, K and L are substitutable factors, as indicated by equation (12.1). Mineral resources enter into the production process in a fixed proportion to gross output, which is represented by the coefficient z , and is technologically determined.

Finally, the production system (12.1)-(12.2) also assumes given values for the length of the working period and the work intensity in the production units. For the analysis of the long run, which is the one that concerns us here, the unit of time would be a long period, say, the decade.

Some of these assumptions will be modified by constructing two more models later on. The laws of thermodynamics (dealing with matter and energy relations) will be introduced in model B; substitution between funds and flows will be discussed in model C. Model B will turned out to be the proper model.

In model A, consider for a moment that mineral resources are redundant factors; therefore, the relevant equation in the production system is equation (12.1). Net output is by definition equal to gross output minus the quantity of goods devoted to the reposition of the stock K . The term “reposition” in this case means the quantity of good B needed to maintain constant the stock K , which implies securing the same stocks and thus the same quantity of service funds, period after period.

Let b represent the coefficient of reposition of K . Hence, b indicates the quantity of good B that is needed per unit of K to maintain it constant, which multiplied by the quantity of K will give us the total quantity of good B needed to replace directly the wear and tear of machines and thus keep the stock of capital K constant period after period. The model assumes effective full employment of machines and men.

The reposition equation for any period T can then be written as

$$\begin{aligned} R(T) &= b K \\ &= r Y^*(T), \quad 0 < r < 1 \end{aligned} \tag{12.3}$$

Therefore, R indicates total reposition, that is, the total quantity of good B that is needed to maintain constant the stock of K . For a given K/Y^* ratio (dynamic equilibrium), the flow of reposition R can be represented as a fixed proportion of the flow of gross output, the coefficient r . Because we are dealing with a production process that is productive, the coefficient r must be less than one.

The flow of net output Y can then be written as

$$\begin{aligned} Y(T) &= Y^*(T) - R(T) \\ &= Y^*(T) - r Y^*(T) \\ &= (1 - r) Y^*(T) \end{aligned} \quad (12.4)$$

This equation shows that the flow of net output Y is a fixed proportion of the flow of gross output Y^* . The use of reposition makes the stock K a *renewable* factor and net output sustainable over time; that is, net output can be repeated period after period, as long as the mineral resources are redundant. Therefore, the net output Y of any period can be allocated to capital accumulation (as physical and human) and to consumption.

Consider that society is endowed with machines and men in quantities K_1 and L_1 , which are now the redundant factors of production. Then the relevant equation in the production system is equation (12.2). The initial stock of mineral resources S_0 will decrease continuously in the production process, even if the *same quantity* of gross output is produced period after period. Therefore, the quantity remaining of the stock of mineral resources at the end of the period T , the term $S(T)$, can be written as

$$\begin{aligned} S(T) &= S_0 - \sum N_j = S_0 - \sum z Y_j^*, j=1, 2, \dots, T \\ &= S_0 - z Y^* T \end{aligned} \quad (12.5)$$

If the quantity of gross output is given (Y^*), the initial stock of mineral resources declines irrevocably over time at the rate of $N=zY^*$ per unit of time. The new stock at period T becomes $S(T)$ according to the number of periods that the production process was repeated. The period at which the stock of mineral resources become depleted can be found by setting $S(T)=0$.

12.3 The Intergenerational Consumption Frontier

Both equations of the production system (12.1)-(12.2) will now be taken into account. Let K_1 and L_1 represent the society's factor endowments of machines and men, which can produce gross output Y_1^* , and which makes the initial stock of mineral resources redundant. This relation will then be introduced into equation (12.5). Thus

$$\begin{aligned} S(T) &= S_0 - z F(K_1, L_1) T \\ &= S_0 - z Y_1^* T \\ &= S_0 - [z/(1-r)] Y_1 T \\ &= S_0 - \mu Y_1 T, \text{ where } \mu = z/(1-r) \end{aligned} \quad (12.6)$$

The depletion rate of the initial stock of mineral resources is now presented in terms of the net output Y_1 . The stock of mineral resources declines irrevocably over time at the rate of $N=\mu Y_1$ per unit of time, where μ represents the technological requirement of mineral resources per unit of net output.

The period at which the stock of mineral resources is eventually depleted can be found by setting equation (12.6) equal to zero, that is, $S_0=\mu Y_1 T$. This equality shows that, given the initial stock of mineral resources and given the technological coefficient, the total output to be ever produced (YT) is fixed. Two properties of the production process then appear: (a) if the net output is fixed, there will be a finite period T' at which the stock is depleted; (b) if the net output is doubled, the number of periods that it can be repeated will be reduced to half or if the net output is reduced to half, the number of periods will double.

Dividing equation (12.6) by μ , the time path of the stock of mineral resources $S(T)$ can be transformed into the time path of net output $Y(T)$. Hence,

$$Y(T) = S_0/\mu - Y_1 T \quad (12.7)$$

The first term on the right-hand side shows the society's endowment in units of good B. As net output Y_1 is produced and repeated over time, the stock of mineral resources declines over time at the rate given by Y_1 .

Equation (12.7) represents the constraints of both funds and flows in the production of net output Y_1 , such that the stock of mineral resources is initially redundant factor. Then as the net output Y_1 is repeated period after period, the stock of mineral resources will decrease continuously and irrevocably, until it is ultimately depleted. By setting $Y(T)=0$, this period is determined; call it $T=T'$. At this period, net output will become zero. Certainly, T' would imply the extinction of the human society. The period at which mineral resources stop being redundant can also be determined easily by setting $Y(T)=Y_1$; call this $T=T^*$. It is clear that $T^*=T'-1$.

Another assumption will now be introduced into the model. Society will not let nature determine the end of its history and thus confronted with the risk of extinction will take actions; in particular, assume society will decide at period T^* (when mineral resources are no longer redundant) to extend the duration of the remaining stock of mineral resources for more than one period by setting the consumption at a lower level. The remaining stock of mineral resources can then be extended over several periods and used at the rate given by the new consumption level until these resources become depleted. This end period, socially determined, will be called T° , such that $T^*<T'<T^\circ$.

Equation (12.7) is represented in Figure 12.1. The horizontal axis measures historical time and the vertical axis net output. Given the stocks of K_1 and L_1 , and also given the level of technology A_1 , the corresponding flow of net output is represented by the segment OA , that is, $Y_1=OA$. The mineral resources constraint is given by the line MV . Then OM units of net output could be produced initially with the given stock of mineral resources; hence, mineral resources are initially redundant. But as OA units of net output are repeated period after period, the stock of mineral resources will decrease until the stock is depleted, which occurs at period T_B (period 6 in Figure 12.1). This is the basic nature of the flow-fund production process, as initially represented by the production system (12.1)-(12.2).

Net output is equal to consumption in Figure 12.1. At period T_B^* (period 5), when mineral resources are no longer redundant, society decides to intervene and extend the duration of the remaining mineral resources by reducing consumption to a fraction of the current consumption level OA. The mineral resources left unused in the final period can then last one more period or several periods until they are eventually depleted, which depends upon the social choice regarding the new level of consumption. If the choice is for maintaining the consumption level, and thus extend for one additional period, then the choice is for point B'; if the fraction is one-half, the extension will be for two periods; if the fraction is one-third, the extension will be for three periods, and so on, which is shown by the curve B'Z, which is an equilateral hyperbola.

The time path of the consumption possibilities may be called the *intergenerational consumption frontier*. It is then represented by the segment AB and the social choice of one, and only one, point on the curve B'Z (thus represented by a discontinuous line). The segment AB is constrained by the stocks of K_1 and L_1 and the segment BZ by the stock of mineral resources S_0 .

Let the social choice on the segment B'Z be point P. Beyond period T_B^* (period 5), the consumption level is given by the segment CP (the level C is one-fourth of OA in Figure 12.1) and will last for four additional periods, until mineral resources become depleted in period T_B° (period 9). The initial stocks of workers and machines now become redundant; the quantity of net output is limited by the available mineral resources. To simplify, assume that the number of workers remains unchanged, even though only a fraction of total workers are needed in production. Some institutional changes will have to be introduced into society to accommodate this separation between production and distribution: although only a fraction of workers are needed in production, all workers will participate in the distribution of total output. The stock of machines will be let to decline for no total repossession is needed.

The distribution of consumption between generations can then be seen in the intergenerational consumption frontier, as shown in Figure 12.1. The consumption level of the present generation (OA for period 1) will be higher than the *average* consumption level of future generations (OA for four generations and OC for four generations). Consequently, there is consumption inequality between generations. This is so even maintaining fixed the consumption level of the current generations. The reason lies in the finite stock of mineral resources, which impede that the consumption level OA could go on forever.

In model A, in conclusion, when the stock of non-renewable natural resources is included in the production process, the consumption level of the current generation cannot be repeated period after period forever. This is just the result of the inevitable depletion of a given stock of exhaustible resources. Moreover, there will be a degree of inequality in the level of consumption between generations: the average consumption level of future generations will necessarily be smaller than that of the current generation.

12.4 Model B: Introducing the Laws of Thermodynamics

Textbooks on environmental economics usually recognize two schools in this new discipline (Hanley, Shogren, and White 2001). The standard economic theory of the environment is based on neoclassical theory and on the first law of thermodynamics. The

other school, called bio-economics, was initiated by Georgescu-Roegen (1971), who introduced the second law of thermodynamics—the entropy law—into the economic process. Both laws of thermodynamics are introduced now into the analysis of the production process, in the production system (12.1)-(12.2) and its derived relations.

So far the effect of consumption on the environment has operated through the continuous decrease in the stock of mineral resources until its ultimate depletion. This effect may be called the *pure depletion effect* of a given stock of non-renewable resources.

The laws of thermodynamics that are of interest in the economic process were put simply by Georgescu-Roegen (1971), which is worth quoting here:

Let us take the case of an old-fashioned railway engine in which heat of the burning coal flows into the boiler and, through the escaping steam, from the boiler into the atmosphere. One obvious result of this process is some mechanical work: the train has moved from one station to another. To wit, the coal has been transformed into ashes. Yet one thing is certain: the total quantity of matter and energy has not been altered. That is the dictate of the Law of the Conservation of Matter and Energy—which is the First Law of Thermodynamics ... At the beginning the chemical energy of the coal is *free*, in the sense that it is available to us for producing some mechanical work. In the process, however, the free energy loses its quality, bit by bit. Ultimately, it always dissipates completely into the whole system where it becomes *bound* energy, that is, energy which we can no longer use for the same purpose. ... In other words, high entropy means a structure in which most or all energy is bound, and low entropy, a structure in which the opposite is true. ... [This is] the Entropy Law, which is the Second Law of Thermodynamics. All it says is that the entropy of the universe (or of an isolated structure) increases constantly ... and irrevocably. We may say that in the universe there is a *continuous* and *irrevocable* qualitative degradation of free into bound energy (pp. 5-6).

The outcome of the production process includes not only good B, but also “bads” because waste is another irrevocable outcome of the production process. This is just the constraint given by the First Law of Thermodynamics: matter and energy can only be rearranged, not destroyed or created.

The First Law has another implication in the production process. The production of material goods consists in the transformation of some materials into others (the flow elements of production) by some agents (the fund elements). Therefore, mineral resources are *essential* elements in the production process in the sense that $N=0$ implies $Y^*=Y_1=0$. This property was already introduced as assumption of the production system (12.1)-(12.2), in which technology is limitational. According to the Second Law of Thermodynamics, waste is transformed into pollution of the biophysical environment. Depletion and pollution constitute the two forms in which the economic process causes degradation of the environment.

The economic process is dependent upon the environment in those two ways: (a) as a source of mineral resources (low entropy) and (b) as a sink for waste (high entropy), which together degrade the environment (Daly, 1996, p. 33). The finite size of Earth imposes limits to both components: it implies a given stock of mineral resources and also a finite capacity to absorb waste, which means that the absorptive capacity of the ecosystem is limited if it is going to be able to continue supporting human life, *as we know it*. The

given stock of mineral resources would not be a problem in the economic process if everything could be recycled, but the entropy law prevents full recycling; waste would not be a problem if the absorptive capacity of the ecosystem were infinite.

In the economic process, therefore, even a *constant* net output flow implies a continuous and irrevocable depletion of mineral resources and pollution of the environment. Therefore, the economic process is a human activity that can also be seen as the transformation of low entropy (mineral resources) into high entropy (waste and pollution). Available matter and available energy can be used only once in the production process. The production process thus implies degradation of free into bound energy.

Both laws of thermodynamics are very much interrelated. As economist Kenneth Boulding stated:

In a closed system, the first law says that all that can happen is rearrangement; the second law says that if rearrangement happens, it is because there is some kind of potential for rearrangement, and as rearrangement goes on, potential is gradually reduced to zero and we get to the point where nothing further can happen (Boulding 1976,p.5).

The production process only rearranges matter and energy, but in doing so the production capacity is qualitatively degraded. Therefore, as the production is repeated period after period, the potential of the production system is continuously and irrevocably degraded. The economic process is not mechanical, but entropic.

How do the laws of thermodynamics affect the intergenerational consumption frontier? Firstly, the effect of waste on the qualitative degradation of the biophysical environment must be taken into account. Waste implies pollution of the environment, including water, air, and soil. Assume now that pollution increases the average global temperature and that this climate change will affect the production process by making it more risky.

Secondly, pollution is an outcome of the production process, but it now has a feedback effect upon the production process because pollution will increase the cost of reposition of machines. Due to the direct damage of pollution upon the physical capital and due to the higher risk of destruction from climate change, a higher depreciation rate will now be required in order to maintain the stock of machines both productive and durable.

As a result, more mineral resources will be required to maintain the same level of net output. Because the flow of pollution *accumulates* in the environment, as the same net output is produced period after period, the feedback effect will increase over time, and thus the technological coefficient of mineral resources required per unit of net output will increase over time; that is, the value of the coefficient μ will increase continuously and irrevocably over time.

The initial assumption on the production process indicated by the system (12.1)-(12.2) will now be modified. For a given stocks of K_1 and L_1 , the flow of gross output Y_1^* will be produced, provided mineral resources will be flowing into the production process in the quantity of N , which now includes the requirements of both direct material inputs and the indirect inputs induced by the stock of pollution (Π). This stock at period T can be written as

$$\Pi(T) = \sum \beta N_j = \beta \sum Y_j^*, j=1, 2, \dots, T \quad (12.8)$$

The coefficient β indicates the pollution rate or the rate of greenhouse gas emissions from using mineral resources in the production process.

Then we can include the feedback of pollution into the production process to determine the total coefficient of mineral resources per unit of net output. Firstly, the reposition costs of machines (R) are now:

$$\begin{aligned} R(T) &= r Y_1^* + r' \Pi(T) = r Y_1^* + r' \beta \sum Y_j^*, j=1, 2, \dots, T \\ &= r Y_1^* + r' \beta \sum Y_1^* T \\ &= (r + r' \beta \sum T) Y_1^* \\ &= \lambda(T) Y_1^*, \text{ where } \lambda(T) = r + r' \beta \sum T \end{aligned} \quad (12.9)$$

In equation (12.9), the first term shows the usual reposition cost, which is equal to the proportion r of total gross output; then the second refers to the costs of reposition due to the pollution effect on machines, which is equal to the proportion (r') of the stock of pollution. Therefore, the coefficient of total reposition per unit of gross output is represented by λ , which increases over time, and as a function of time T is represented by $\lambda(T)$.

The new relation between net output and gross output then becomes

$$\begin{aligned} Y(T) &= Y^*(T) - R(T) = Y_1^* - \lambda(T) Y_1^* \\ &= Y_1^* [1 - \lambda(T)], \text{ where } [1 - \lambda] > 0, \text{ and } \lambda'(T) > 0 \end{aligned} \quad (12.10)$$

The quantity of mineral resources required per unit of net output is now

$$\begin{aligned} N(T) &= \sum Y_1^* \\ &= (z / [1 - \lambda(T)]) Y_1^* T \\ &= \varepsilon(T) Y_1^*, \text{ where } \varepsilon(T) = z / [1 - \lambda(T)], \varepsilon'(T) > 0, \text{ and } \varepsilon(0) = z / (1 - r) = \mu \end{aligned} \quad (12.11)$$

The coefficient ε represents the quantity of mineral resources per unit of net output, the value of which includes the feed-back effect of pollution upon the production process; moreover, the value of this coefficient increases over time due to the cumulative effect of production upon pollution.

In order to derive the time path of the consumption possibilities frontier, we have to re-write equation (12.6), in which the stocks K_1 and L_1 can produce gross output Y_1^* , taking into account the new relations established in equation (12.11). Then

$$\begin{aligned} S(T) &= S_0 - \sum N_j = S_0 - z \sum Y_j^* = S_0 - z Y_1^* T \\ &= S_0 - (z / [1 - \lambda(T)]) Y_1^* T \\ &= S_0 - \varepsilon(T) Y_1^* T \end{aligned} \quad (12.12)$$

$$\begin{aligned} Y(T) &= [S_0 / \varepsilon(T)] - \sum Y_j, j=1, 2, \dots, T \\ &= [S_0 / \varepsilon(T)] - Y_1 T \end{aligned} \quad (12.13)$$

Equation (12.12) shows the time path of the stock of mineral resources, which at any time T is equal to the initial stock S_0 minus the quantity used up to that period. The quantity of

net output Y_1 is determined by the funds (the stocks of machines and workers) and is integrated into the constraint given by the stock of mineral resources.

Equation (12.13) is just the result of dividing equation (12.12) by the coefficient ε , and shows the time path of net output determined by the constraints of mineral resources. It is clear that as T increases, the requirement of mineral resources per unit of net output (the coefficient ε) increases, which implies a continuous shift downwards of the intercept of the frontier. The time path of output that is determined by the mineral resources constraint is not linear, but a convex curve. Thus the same net output will lead to the depletion of the mineral resources sooner compared to the previous case, when the pollution feedback was ignored in the production process.

The entropic production process summarized by equation (12.3) is represented in Figure 12.2. The depletion effect is shown in the panel (a). The straight line MV assumes a constant technological coefficient of mineral resources required per unit of net output (as in Figure 12.1). Since this coefficient increases over time, the initial straight line MW will continuously be shifted inward and will become a convex curve, which will end to the left of point V , and will cut the segment AB before point B , at point E . The entropic production process implies a more rapid depletion rate of mineral resources, now showed by the curve MW .

The social intervention period takes place when the horizontal line AB is cut by the curve MW , at point E , which occurs at period T_d^* . Beyond this period, with the remaining mineral resources, the set of future consumption possibilities will be given by the curve $E'X$. Then social choice can determine one, and only one, particular point along this curve. The intergenerational consumption frontier is now given by the segment AE and the social choice of a point on the curve $E'X$. Clearly, this consumption path is more limited than the comparable path given by the segment AB and a point on the segment $B'Z$, shown in Figure 12.1. The difference is due to the effect of the entropy law.

Figure 12.2, panel (b), depicts separately the pollution consequences of the economic process. As the same quantity of net output is repeated over time, the environment accumulates increasing stocks of waste and pollution, which is represented by the curve $O'G$. Hence the curve is rapidly rising up to the period T_d^* , when mineral resources become scarce. Beyond this period, with social intervention, the stock of pollution still grows but at slower rates, and then follows the path FG' , rather than FG ; that is, as the consumption level falls, the rate of increase of pollution also falls, but the stock increases continuously and irrevocably, as the entropy law says. Society can modify the *rate* of degradation, but not the degradation itself.

Figure 12.2 thus shows the two laws of thermodynamics in action. These laws impose constraints to the economic process through depletion and pollution. These effects are inter-related, as suggested in the figure. The depletion effect of mineral resources sets a time limit to the production of a given net output: output OA can be repeated until period T_d^* . Pollution will have the same property. There is a limited capacity of the ecosystem to absorb waste if it is going to maintain its capacity to continue supporting human life, as we know it. Let this limited capacity be represented as a threshold, given by the level $O'C$, which occurs at period T_p^* . If pollution in the atmosphere reaches values beyond this threshold, the preservation of human life, as we know it, could not continue. Some

qualitative changes and adaptation in human life would then take place; for example, low oxygen availability in the air could lead us to a kind of anaerobic human life.

Depletion and pollution resulting from the economic process will impose different threshold periods for the existence of human species, depending on which of the two periods comes first. In Figure 12.2, for instance, our model assumes that the pollution threshold (T_p^*) will come sooner than that of depletion (T_d^*). The relevant constraint of the environment is, in this case, the capacity of the ecological system to support human society, not the depletion of mineral resources. This ecological capacity is the ultimate element of scarcity in the economic process. Everything can be produced or substituted, except the ecological capacity. The model says that we humans cannot supply ourselves with another ecological environment. Therefore, the model predicts that intergenerational consumption frontier shown in panel (a) will not attain the entire time path, but only up to AE”.

The model assumes that the human society will take actions when confronted with the risk of its extinction. In this case the human society would have to move to another age via technological and institutional innovations. The case shown in Figure 12.2 says that the age of mineral resources will be abandoned before the mineral resources have been exhausted. In the history of human society, the stone-age was abandoned not because stones became scarce.

Figure 12.2 indicates clearly that the ecological conflict of mankind holds even if the level of consumption remains fixed, that is, even if we had a zero-growth society, in output and in population. Certainly, the conflict will be more acute if society embarks in a process of economic growth, as will be shown below.

The role of renewable natural resources in the economic process has been ignored in this entropic model. The implicit assumption has been that these resources were redundant, which will now be revised. For this purpose, two sources of energy must now be distinguished in the production process: (1) the finite stocks of mineral resources in the Earth’s crust, which is exhaustible; (2) the sun’s stock of energy, from which the flow of solar radiation comes to Earth and is the source of energy for the existence of renewable natural resources, such as forestry and fisheries.

The Earth is a closed thermodynamic system in the sense that it does not exchange matter with outer space, but only energy from the sun (Baumgärtner 2004, p. 320). Then the scarcity of renewable natural resources comes from the Earth’s limited size as a catching net of the solar energy. Agricultural soil, for instance, is of limited size; in addition, it is also subject to degradation due to erosion; thus soil belongs to the category of non-renewable resources.

Fisheries, forestry, and other biological resources may however be subject to depletion if the rate of biological reposition is smaller than the rate of exploitation by humans. When *renewable* natural resources are not *renewed*, they will be subject to depletion, just as in the case of mineral resources. If this is the case, renewable natural resources may be considered already included in the coefficients that determine the intergenerational consumption frontier in the model. Those renewable natural resources that are in fact renewed may be considered as redundant factors in the model and may thus be ignored in the analysis.

Solar energy is an absolute redundant factor of production. Then it can be introduced as a horizontal line in Figure 12.2(a), starting from a point above point M. Under this assumption, the consumption path shown in the graph may still represent the intergenerational consumption frontier, which is now determined not only by the constraint of non-renewable resources, but also by the constraint of those renewable resources that the human production activity has transformed into non-renewable.

12.5 Model C: Introducing Substitutability between Funds and Flows

Standard economics has another set of assumptions about the production process. They are summarized in the popular concept of *production function*, which is usually represented as follows:

$$Y = F(K, L, N) \quad (12.14)$$

Thus, standard economics assumes that the quantity of output produced depends upon the stocks of machines, workers, and natural resources, so that these factors of production are all substitutable (Solow 1974, p.34). This innocent equation says that the standard production theory assumes implicitly that the three factors are substitutable in the production process; hence, net output could be produced with machines and workers alone; so net output could be produced forever. Mineral resources are *non-essential* factors of production. Note the difference with the flow-fund approach, which was represented as a *production system*, equations (12.1)-(12.2), rather than as one-equation production function.

A consequence of the standard economics assumption about the production process is that the production of a given net output can go on forever. Therefore, output growth can also go on forever. There are no limits to the production of goods. This view can be summarized as follows:

It is now generally accepted that the limited supply of non-renewable resources does not necessarily imply a limit to growth. In particular, the neoclassical theory gives rise to three main possibilities: (i) substitution of the resource by other inputs, such as capital; (ii) improve of the resource efficiency; and (iii) development of backstop technologies. However, without any technical change, none of these outcomes will balance the resource exhaustion and continue to sustain some positive growth in the long run (Lafforgue 2008, p. 541).

According to this view, a way to introduce substitution between machines and mineral resources would be by assuming that the technological coefficient of mineral resources per unit of net output can fall as the stocks of machines increase. This substitution would be induced by changes in the relative prices of minerals, that is, as mineral resources become more expensive.

Even accepting the possibility of substitution, the question is, where would the additional quantity of machines come from? It would have to be produced and then more mineral resources will be used up in its production. Then the net effect of substitution on the saving of mineral resources would be smaller than what the pure substitution effect

implies. (Wind mills can substitute oil in generating energy, but the posts of wind mills produced in factories would need minerals and other inputs.) In addition, the net output is a material good, which cannot be totally dematerialized, for that would go against the first law of thermodynamics, which sets a limit to the substitution possibilities.

In Figure 12.1, if a quantity of capital can substitute mineral resources, then the line MV would be shifted outward, to another line (say to line $M''V''$, not drawn). But producing that quantity of capital would require mineral resources and would also imply reposition costs in terms of mineral resources. So the net effect of substituting minerals would be smaller than the initial effect (a change from line MV to, say, line $M'V'$, which would lie below line $M''V''$). Assume that the net effect is positive. Then the intergenerational consumption frontier would be shifted outward. Consequently, period T^* would be expanded, but it would still be finite. More substitution could proceed, but a production limit would be reached sooner or later. Let the line MV represent this limit in substitution possibilities and the model will have captured the substitution effect.

In sum, in the entropic production process, substitution between fund and flow factors is possible, but to a certain extent only. This is due to the assumption that mineral resources are *essential* factors of production, which is consistent with the laws of thermodynamics. However, these substitution effects will not eliminate the existence of the intergenerational consumption frontier. Even with substitution, as long as a given net output is repeated period after period, mineral resources will eventually become scarce and depleted and pollution will increase. Therefore, as long as mineral resources are essential factors of production in the production process, this conclusion will hold true.

In short, there is no need to change the conclusions reached so far in the entropic model B. Model B is the appropriate model to study the interactions between the economic process and the environment.

12.6 Changes in the Intergenerational Consumption Frontier

The intergenerational consumption frontier has been constructed under several givens. The exogenous variables of the model B include technology, and the endowments of machines, workers, and mineral resources. It is time to analyze the effect of changes in these exogenous variables upon the intergenerational consumption frontier.

An exogenous increase in the stocks of machines and workers, together with technological change that is incorporated in the new investments in physical capital and human capital, will increase the current flow of gross output and net output; hence the consumption level of the current generation will also increase. But then the rate of depletion of the given stock of mineral resources will also increase, which will in turn increase the pollution rate. In Figure 12.2, higher stocks of K and L will shift the level of the intergenerational consumption frontier OA upward, which implies an inward shift of the depletion curve MW and an upward shift of the pollution curve $O'G$. Thus the critical periods T' and T^* will occur sooner.

Another consequence is that the degree of intergenerational inequality will be higher: the consumption level of the present generation will increase, but the average consumption level of the future generations will fall. In other words, economic growth

implies an increase in the intergenerational inequality. Therefore, the only choice society has is the distribution of the consumption level, and the corresponding distribution of non-renewable resources, between generations. A higher consumption level allocated to the current generation will imply more mineral resources allocated to the current generation and, consequently, less amount of mineral resources will be left for the future generations, which in turn implies a lower total consumption level for them.

Consider now an exogenous technological change that generates a decrease in the initial coefficient of mineral resources per unit of net output, the coefficient ε . This coefficient is determined by the initial coefficients z and λ , as shown in equation (12.11). A reduction in the value of this technological requirement is equivalent to an increase in the initial stock of mineral resources. This is a mineral resource-saving technological change. Therefore, the intergenerational consumption frontier will be shifted outward and, consequently, the pollution curve (which also depends upon coefficient ε) will be shifted downwards.

These effects can also be visualized in Figure 12.2. With new technologies that save mineral resources per unit of net output, the intercept of the mineral resources constraint curve will move from M to another point above it; thus the curve MEW will be shifted outwards; consequently, the intergenerational consumption frontier will also be shifted outwards and the pollution curve $O'G$ will be shifted downward. From equations (12.8) and (12.10), we can see that the reason for this shift is that the curve $O'G$ is determined by the flow of net output ($Y_1=OA$), which does not change, and also by the technological coefficients, which do change. As a result, the critical periods T' and T^* will not be eliminated, just will occur later.

It is still true, however, that the current consumption level cannot be repeated period after period forever; consequently, technological progress cannot eliminate the existence of the intergenerational consumption frontier; it can only move the frontier to another level. At each new level of technology, there will be a new intergenerational consumption frontier; moreover, this new frontier will reduce the degree of inequality between generations. This is assuming that technological change is cost-free. Taking into account the cost of mineral resources in research and development (R&D), the net effect would be smaller.

Could technological change be so strong in saving mineral resources that a given consumption level would be maintained forever? Could technological change eliminate the constraints imposed by the two laws of thermodynamics?

Assume technological change is endogenous and is cost-free. One could imagine that if mineral resources are depleted by half in a given period of production, technological change could immediately occur and reduce the technological coefficient of minerals per unit of net output also by half, which is equivalent to increasing mineral resources by double. The consequence would be that the stock of mineral resources remained constant over time, that is, mineral resources would have now become *renewable* natural resource. Then the consumption level OA could be repeated forever in Figure 12.2. Along this horizontal line, machines, workers, and minerals would all become renewable resources thanks to technological change.

However, the panel (b) of Figure 12.2 needs to be taken into consideration. The pollution effect will continue irrevocably. Mineral resources will be used up to produce Y_1

in the first period; although the stock of mineral resources is economically recovered through technological change, the amount of mineral resources used up will have generated pollution. In the next period, net output will be repeated and mineral resources will be used up; and although the stock of mineral resources is economically recovered, the pollution effect will have taken place and will have accumulated for two periods, and so on. The curve O'G will become linear. Then pollution, not depletion, would be the limiting factor of the economic process. Technological change would now have to eliminate the stock of pollution in order to have a non-entropic production process. Technological change would have to solve two problems: depletion and pollution. Under the most favorable scenario, it is unlikely that technological change can eliminate the laws of thermodynamics.

Economic growth combined with mineral resource-saving technological change seem to have an ambiguous effect on the threshold periods, T' and T^* . The growth effect reduces those threshold values, but the technology effect extends them. However, given the argument presented above about the limits of technological change, the economic growth effect would tend to prevail. If the human society chooses economic growth, then the survival of the human society, as we know it, would be shorter.

12.7 Economic Growth as Evolutionary Process

The nature of equilibrium in the entropic model B needs to be clearly understood. As we have seen, even for a constant consumption level, the economic process goes through quantitative and qualitative changes. The mineral resources tend to be depleted and the environment is polluted. The capacity of the environment to support human life, as we know it, is limited. The human society, as any other biological species, runs the risk of its extinction. In the struggle against that destiny, the human society would tend to adapt and some institutional changes would have to occur to adapt to that situation. The entropic economic process can then be seen as an evolutionary process, in which both quantitative and qualitative changes takes place.

To be sure, the entropic economic process does not operate under static equilibrium. The standard definition of static equilibrium says: the value of the endogenous variable (the quantity of net output produced) will remain fixed and be repeated period after period as long as the values of the exogenous variables remain fixed. In the entropic model, however, the stock of mineral resources (an exogenous variable) falls in every period and pollution accumulates in every period.

The entropic model does not operate under a dynamic equilibrium either. The standard definition of dynamic equilibrium says: the value of the endogenous variable (quantity of net output produced) will move along a particular trajectory over time as long as the exogenous variables remain fixed. Only quantitative changes will take place. However, quantitative and qualitative changes occur in the entropic process, which is underlying the path of the intergenerational consumption frontier.

When the laws of thermodynamics are included into the production process, then some surprising properties of the economic process appear. Even static equilibrium is unviable. Even if the economic process is repeated *at the same level of net output* period after period, the society's capacity to reproduce itself at that level of consumption forever is unviable. Certainly, the dynamic equilibrium of output growth will also be unviable. As net

output increases over time, both the rate of depletion and the rate of pollution will increase. Economic growth cannot go on forever. This conclusion can be visualized with the help of Figure 12.2.

How does economic growth proceed under an entropic economic process? Growth of net output requires capital accumulation. Assume that technological change is incorporated in the new capital. As a result of capital accumulation, the average net output for the current generation will increase. At this new output flow, mineral resources would be depleted at higher rates and pollution will increase also at a higher rates; therefore, the critical values T^* and T' will thus occur sooner than without economic growth (as shown above in Section 12.6, as the combined effect of capital accumulation with technological change).

The continuous increase in total net output over time would imply not only a time path of depletion of the stock of mineral resources but also a time path of pollution, which has a threshold of tolerable pollution level. Then the constraint in the economic growth process would be pollution, not depletion. Beyond this threshold, the environment would be unable to support human life, as we know it. New forms of human life would appear in new ecological systems, possibly a more anaerobic human life. At this threshold period society will then take actions to adjust and adapt to the new situation. New social institutions will be introduced in the workings of society, to cope with the necessary separation between production and distribution, as shown above.

The evolutionary process of economic growth is shown in Figure 12.3. The vertical axis measures output per worker and the horizontal axis measures historical time T . Consider the case in which mineral resources are initially redundant factors of production; hence, production is limited by the accumulation of capital and the growth rate of population and workers. Curve DR represents the economic growth frontier in terms of output per worker, in which the exogenous variables are the initial inequality in the world society (δ) and the rate of progress in the technology that is labor augmenting (τ_1).

This relation is similar to equation (11.17), presented in Chapter 11, which showed the growth frontier for each type of society. Now we are dealing with the capitalist system as a whole, equivalent to the world society in this model, and then some changes are in order: initial inequality refers to the entire capitalist system and g is replaced by τ_1 because two types of technologies are now in operation. Furthermore, and just for the sake of simplicity, assume that the third exogenous factor, the world savings ratio is endogenous and depends positively upon the degree of the initial inequality. Then for the world society, the time path of output per worker can be written as:

$$y(T) = F(T; \delta_0, \tau_1), F_1 > 0, F_2 < 0, F_3 > 0 \quad (12.15)$$

This equation assumes that output per worker increases over time at a given growth rate. The exogenous variable initial inequality has a level effect, while the rate of technological progress has a growth effect.

Now introduce the restrictions imposed by the two laws of thermodynamics. Given stock of mineral resources, as economic growth proceeds, the stock of mineral resources will decline, and output per worker that is constrained by mineral resources will fall over time, as illustrated by the curve BN. Assume now that this time path is given by the dynamic equilibrium output per worker studied in Chapter 11, with the following figures as

illustration: total output grows (5% per year) at the same rate as capital accumulation (5%) and at the same combined rate of workers (2%) and technological change (3%); hence, output per worker grows at 3%, which is equal to the growth rate of technological progress. In Figure 12.3, the curve DR shows a given growth rate, say at 3% per year. Then mineral resources will be used up at a very rapid growth rate.

The path of mineral resources depletion is determined by three exogenous variables: first, the growth rate of the progress in the technology that is mineral resources saving (τ_2), which implies the reduction of the technological coefficient z contained in the variable ε ; second, the economic growth of output per worker, which is determined by the initial inequality and the rate of technological progress in the labor augmenting technology.

The evolutionary model predicts that as the economic growth proceeds, pollution will increase. This is shown in Figure 12.4. The available measures of CO₂ concentrations in the world atmosphere for the last one thousand years shows a rapid increase starting around the beginning of year 1800, when both the industrial revolution and the beginning of capitalist growth started.

From equations (12.8) and (12.10), we can derive the path of pollution in terms of net output, as follows:

$$\Pi(T) = \sum \beta N_j = \beta \sum z Y_j^* = \beta \varepsilon(T) \sum Y_j, \quad j=1, 2, \dots, T \quad (12.16)$$

This equation shows that the accumulation of pollution in period T is determined by the time path of total net output. Because there is growth, the level of net output in period T will be the largest over time up to that period, and so will net output per worker.

Given the threshold of tolerable pollution (Π^*), equation (12.16) will determine the limit to the time path of net output, which in turn implies an end period (T_p^*), which in turn implies a maximum level of total output (Y^*) and also a maximum level of output per worker (y^*). This latter value is shown by point H in Figure 12.3. The exogenous variables that determine the maximum level of output per capita include those exogenous variables that determine the economic growth rate of y (variables δ and τ_1) and the exogenous variable τ_2 , which lies behind ε .

The limits to economic growth can now be seen in Figure 12.3. As economic growth proceeds, mineral resources will become scarce at point E, where curves DR and BN cross each other, which occurs at period T_d^* . At the same time, as economic growth proceeds, pollution will increase, and the tolerable threshold value will imply a limit to output per worker equal to the horizontal line HM, which crosses the curve DR at point S, which will occur in period T_p^* . The limit to economic growth will be given either by depletion or by pollution.

The incentives of capitalist and governments will be to give priority to strategies that avoid the depletion of minerals. Technological change, discovery of new deposits, ways to reduce the Ricardian diminishing returns in the exploitation of lower quality deposits, and so on. Economic growth cannot be stopped. Incentives for fighting pollution are not as strong. For one thing, it is a public good; actually a public bad. Therefore, the model predicts that pollution will set the limit to economic growth. This prediction is shown in Figure 12.3.

The exogenous variables of the evolutionary model are three, but for the time being ignore the variable initial inequality. Then the relevant exogenous variables are given by the rates of progress of technological progress. The higher the rate of technological progress that is labor augmenting, the higher the rate of economic growth will be, and the curve DR will be shifted upward and thus the curve BN will be shifted inward and curve HM upward. Points E and S will be shifted inward. On the other hand, the higher the technological progress that is mineral saving, the lower the rate of depletion of mineral resources, and the curve BN will be shifted outward and HM upward, which implies that points E and S will also be shifted outward. In any event, the end period of the economic process remains to be finite.

In sum, economic growth cannot go on forever in the evolutionary model, even in the event of technological progress. *According to the entropic model, there is no such thing as sustainable economic growth.*

The mechanical production process, or the circular flow equilibrium, presented in most standard economics textbooks assumes the laws of mechanics, that is, it makes abstraction of the laws of thermodynamics. Standard economics assumes static and dynamic processes; it views the economic process as a mechanical process, in which qualitative changes are ignored; hence, according to this view, economic growth can go on forever.

By contrast, the entropic model assumes evolutionary dynamics in the economic process. Indeed, following Georgescu-Roegen (1971), the economic growth process depicted in Figure 12.3 refers to an evolutionary dynamics, in which *Time T* is historical time (with past, present, and future). This is contrary to Figure 11.6, in which *time t* refers to mechanical time: economic growth moves in the same way irrespective of when the event occurs in historical time (just like a pendulum movement, which is invariable with respect to historical time).

On the other hand, the neoclassical growth models presented in popular textbooks predict that indeed economic growth can proceed forever, in which the role of non-renewable natural resources is ignored (cf. Barro and Sala-i-Martin, 2004). There are some neoclassical models that deal with exhaustible natural resources and its optimal rate of extraction (the Hotelling rule); fewer models deal with the problem of pollution, which is basically treated as a problem of externalities, and thus amenable to solution via Pigouvian taxes (Grimaud and Rouge 2005). The neoclassical models that include natural resources in the economic process are still mechanical; the qualitative consequences of economic growth upon the environment (via the entropy law) are ignored. As a researcher said, “[neoclassical theory assumes that] on the whole thermodynamic constraints are simply irrelevant for economics” (Baumgärtner 2004, p. 308).

12.8 Growth and Quality of Life

Economic growth is, by definition, the continuous increase in the *quantity* of goods per worker or per person produced in society. Economic growth implies higher levels of quantity of goods consumed per person. The higher quantity of goods consumed is often seen as equivalent to the higher *quality* of life in society. This may be true, but under certain conditions only. As will be shown now, this is not true under the current conditions

in which growth, inequality, and environment are related. A distinction must then be made between “average quantity of goods consumed” and “quality of life.”

High quality of life society could be defined as a society in which there is high consumption level, highly equalitarian consumption, and low risk on average consumption. The factors of risk include social disorder, shocks on resource endowments, and shocks on health status. In a view of society that is more comprehensive, the concept of high quality of life should include future generations.

The relation between growth and quality of life, so defined, can easily be established from the theoretical results shown in this chapter. As economic growth proceeds, income inequality either increases or remains constant, but never falls, which implies a higher degree of social disorder. As economic growth proceeds, climate change will take place and shocks on resource endowments will occur more often. Finally, as economic growth proceeds, pollution will increase, which will affect negatively the health status of people.

Therefore, there is an ambiguous relation between economic growth and quality of life. Higher output per worker increases average consumption, but economic growth has negative side effects on the other components of quality of life. The standard view is that growth implies higher quality of life because the side effects are ignored. This relation can be visualized in Figure 12.3: some aspects of the quality of life decrease along the segment DS of the economic growth frontier DR.

As economic growth proceeds, the side effects will gain in importance relative to the direct effect. The reason is that one unit of net output has increasing costs in terms of social disorder, climate change, and pollution. The latter two originate in the effect of the Entropy Law upon the economic process. Therefore, economic growth can initially have a positive net effect on quality of life, but will start having significant negative side effect upon quality of life at some point in the growth process. Economic growth will ultimately have a net negative effect upon the quality of life of society, which will reach its highest point when human society, as we know it, is put to risk of survival. Then there will be an inverted-U shape curve connecting economic growth and quality of life.

Figure 12.5 shows the inverted-U shape relationship that is predicted by the evolutionary model. Before the level y^* , economic growth has a positive net effect upon quality of life; however, beyond this level, the net effect is negative. Due to the Entropy Law it is inevitable that the curve must ultimately slope downward; that is, sooner or later, economic growth will have net negative effect upon quality of life, and eventually will reach the value of zero when the threshold of human survival is reached.

12.9 Empirical Evidence

How does one refute an evolutionary model? The problem was discussed and solved in Chapter 1, which presents the epistemology of this study. The prediction of the evolutionary model about the existence of a threshold value for the future is clearly unfalsifiable. What is falsifiable is the dynamic trajectory of the variables moving towards the threshold value that the model is able to predict.

The entropic model predicts that economic growth contributes to the degradation of the biophysical environment, which eventually will end in an environmental and economic collapse in the future. The prediction of the trajectory is in accord with the observed correlation between growth and pollution in the long run, as indicated as Fact 8, Chapter 2, and in Figure 12.4 above.

The international literature provides additional empirical evidence that tends to give more support to that prediction. The empirical studies show that water, air, and soil pollution is increasing and that they are the result of human activity (Muller 2008). A study by FAO (2005) finds high rates of deforestation at the world scale and shows great concern with these results.

The projections presented in the famous book *Limits to Growth* (Meadows et al 1972) indicated that global economic collapse would occur around year 2030 if growth continued in the same manner, a *ceteris paribus* condition. This collapse referred basically to a fall in the production of food per capita and then to a downturn of population growth. Australian physicist Graham Turner has recently made calculations for the period 1970-2000 and found that the trajectories forecast in that book for this period match very closely the facts (Strauss 2012).

On climate change, the increase in the average temperature of the planet since around 1850 is a fact (IPCC 2007). This period coincides with the industrial revolution and the beginning of capitalist economic growth; it also coincides with the path in CO₂ concentrations shown in Figure 12.4.

Whether the global warming, and the climate change associated to it, is endogenous or exogenous in the economic process is still under scientific debate. The factors affecting climate change can be seen in three steps. First, human fossil-fuel burning causes emissions of greenhouse gases, such as CO₂ concentrations in the air to rise, which has accelerated since the industrial revolution period, as shown in Figure 12.4; second, CO₂ is a greenhouse gas; third, the green-house effect in turn increases average global temperature. The first two are accepted by scientists, but the third is under debate.

For some scientists, the emission of greenhouse gasses implies global warming, which results in climate change; that is, production implies waste and pollution, which results in climate change (Aeschbach-Hertig 2007). Climate change is thus endogenous to the production process. For others, climate change is exogenous to the economic process; it is mainly caused by the natural variations in solar activity (Chilingar, Sorokhtin, and Khilyuk 2008, p. 1572). Yet for others, although it is a complex problem and it is hard to separate the relative significant of each effect with certainty, the conclusion is that part of climate change is endogenous, but unlikely to be the largest part (IPCC 2007, cited in Muller 2008, p. 254).

It should be clear that the entropic model predicts that economic growth is bad for the environment. This is independent of whether the climate change is endogenous or exogenous. If it is endogenous, as the IPCC report indicates, then the negative effect of economic growth upon the environment will be much more important in the entropic model. Indeed, there are empirical studies that show the negative impact of climate change on total average output and its variability. In the case of Latin America, a set of stylized facts are reported by Galindo and Samaniego (2010).

On quality of life, the prediction of the model is about a particular trajectory (the inverted-U shape curve), which includes the present and the future. Observed facts can be used to falsify the model, although the trajectories for the future are unobservable. The available evidence, which is still scarce, tends to be consistent with the empirical prediction: economic growth negative side effects upon quality of life in society operate through pollution, climate change, and social disorder associated to high degree of inequality. The latter has already been corroborated; so the effects through the first two will be presented now.

According to the standard view, infant and child mortality rates, along with life expectancy at birth, are the best indicators of quality of life. These indicators show a significant progress in quality of life in the process of economic growth in both the First World and the Third World. Thus in the period 1970-2004, life expectancy jumped from 72 to 78 years in the First World and from 56 to 65 years in the Third World; infant mortality rates fell from 22 to 5 *per thousand* in the First World and from 108 to 61 in the Third World in the same period (UNDP 2004, Table 9, p.171).

But these indicators refer mostly to “quantity of life,” not to “quality of life” of the survivors. Quality of life in the sense of long life with good health would require data on morbidity rates in the process of economic growth, which are unavailable. If the quantity of life were deflated by the period living in bad health, may be the progress would not like as bright. Similar case appears with infant mortality. The survivors may suffer from malnutrition and their quality of life would thus be low.

The pieces of available evidence tend to show a negative relationship between health and pollution, as predicted by the evolutionary model. Thus the World Health Report of 2004 deals with the environmental health impact and shows the existence of a significant effect: 24% of the global disease burden (healthy life years lost) and 23% of all deaths (premature mortality) can be attributed to environmental factors; moreover, of the 102 mayor diseases, environmental risk factors contributed to disease burden in 85 categories (WHO 2006). Unfortunately, trends of these rates over time are unavailable.

Many local studies also tend to show a negative effect of air pollution upon human health. A study shows negative relations between air pollution and mortality rates for the 88 largest U.S. cities for the period 1987-1994 (Dominici et al 2002). Another found that combustion emissions cause a large number of premature deaths per year in the UK (Yim and Barret 2012). Other studies report the negative effect of air pollution upon asthma in California, USA (Mc Connell et al 2010), upon students health and academic performance in Michigan, USA (Mohai et al 2011), upon children respiratory allergies in USA (Parker Akinbami and Woodruff 2009), upon respiratory illness in infants in USA (Sheffield et al 2011), upon respiratory illness in infants in the Czech Republic (Hertz-Picciotto et al 2007), and upon respiratory health of the elder in India (Mukhopadhyaya and Forssell 2005).

On the prediction of the evolutionary model about the fate of the human society, it is hardly a falsifiable proposition to say that there will be an end of the human society, as we know it, because the future is unobservable. A similar problem of prediction over the future is found in physics. There will be a big crunch of the universe in the future, which is derived from Einstein’s general theory of relativity and the laws of gravity: “gravity is always attractive and implies that the universe must be either expanding or

contracting...there must have been a state of infinite density in the past, the big bang, which would have been an effective beginning of time. Similarly,...there must be another state of infinite density in the future, the big crunch, which would be an end of time” (Hawking 1996, pp. 232).

Gravity theory has explained basic facts of the physical world, but the future big crunch hypothesis is unfalsifiable. Similarly, the evolutionary model of the unified economic theory has also explained several facts of the social world, in the sense that observed facts are consistent with its empirical predictions, but the future end of the human species is unfalsifiable. These two theories are able to explain the observed facts of the real world—the available data cannot refute their predictions—and can thus be accepted provisionally, although their specific predictions about the future are unfalsifiable today.

12.10 Conclusions

This chapter has presented a theoretical model that includes the laws of thermodynamics in the production process. This model is evolutionary. It predicts Fact 8, the last fact of the empirical regularities of capitalism listed in Chapter 2: economic growth implies a continuous and irrevocable degradation of the bio-physical environment in the form of depletion of non-renewable resources and pollution.

According to the evolutionary model, there are limits to economic growth. Economic growth cannot go on forever; therefore, there is no such thing as *sustainable economic growth*.

Economic growth in the context of environmental distress has side effects that are unfavorable to quality of life. Pollution and climate change increase the risk in consumption, life, and health of people. These predictions are consistent with the available empirical evidence. Additionally, the evolutionary model predicts that economic growth increases the consumption inequality between the current generation and future generations.

The evolutionary model predicts that pollution, not depletion, will set the limits to the existence of human society, as we know it. This would occur before mineral resources have been exhausted. This is similar to that period in the history of human society when the stone-age was abandoned before stones were exhausted. In our time, the oil-age could be abandoned before oil stocks are depleted. Another age in human society would have to appear.

Finally, with this chapter the construction of the unified theory of the capitalist system has been completed. The conclusion is that the eight facts about the empirical regularities of production and distribution in the capitalist system have been explained by the unified theory. Therefore, there is no reason to reject the unified theory and we can accept it at this stage of our research, although provisionally, until new empirical evidence or better theory appear. The policy implications of this valid theory can then be analyzed now. To that end the next chapter will summarize the nature of the unified theory.

Figure 12.1. The intergenerational consumption frontier. (Vertical axis measures consumption levels, $OM=120$ and $OA=20$ units; horizontal axis measures generations as time intervals.)

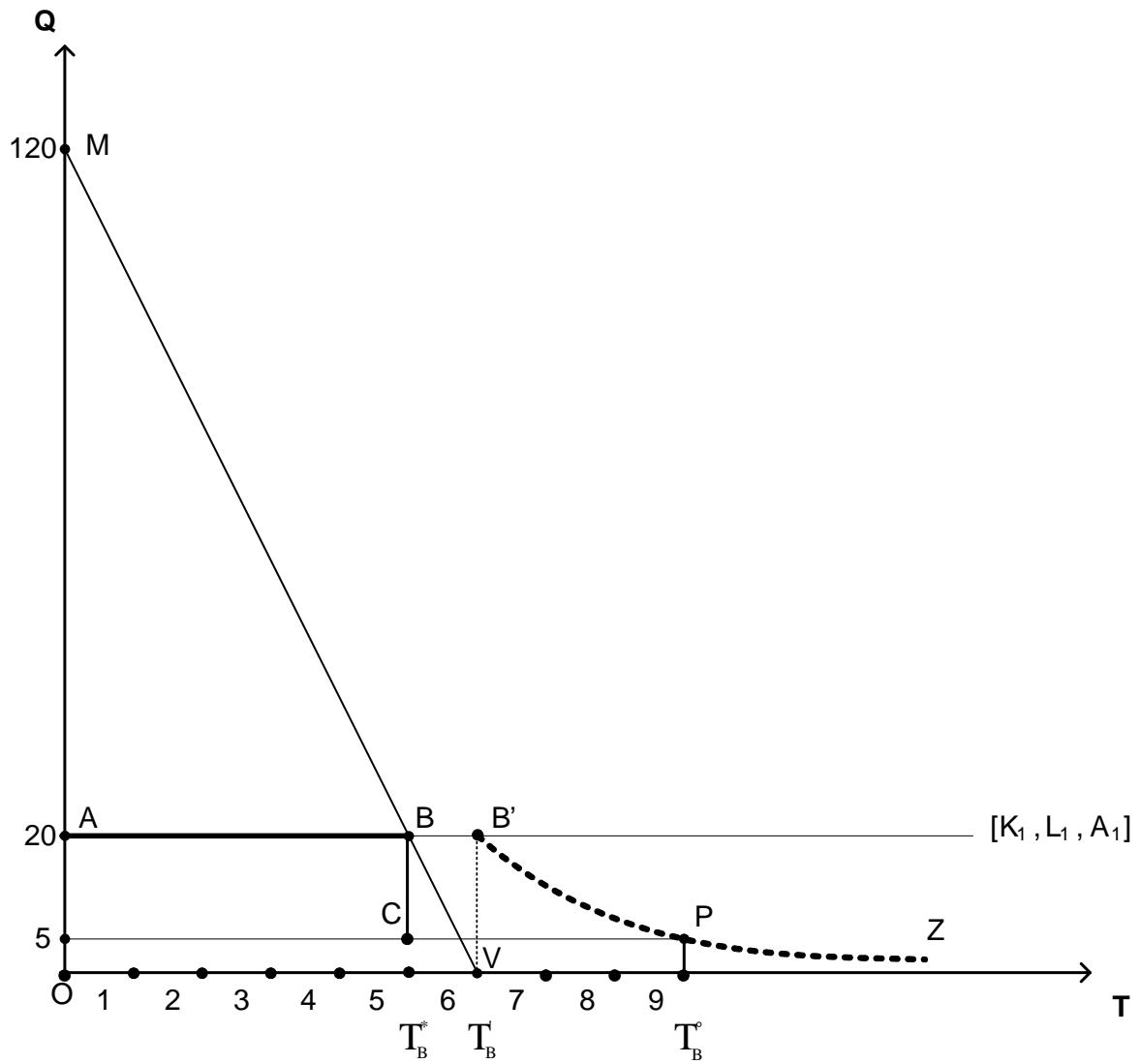


Figure 12.2. Depletion and pollution in the economic process

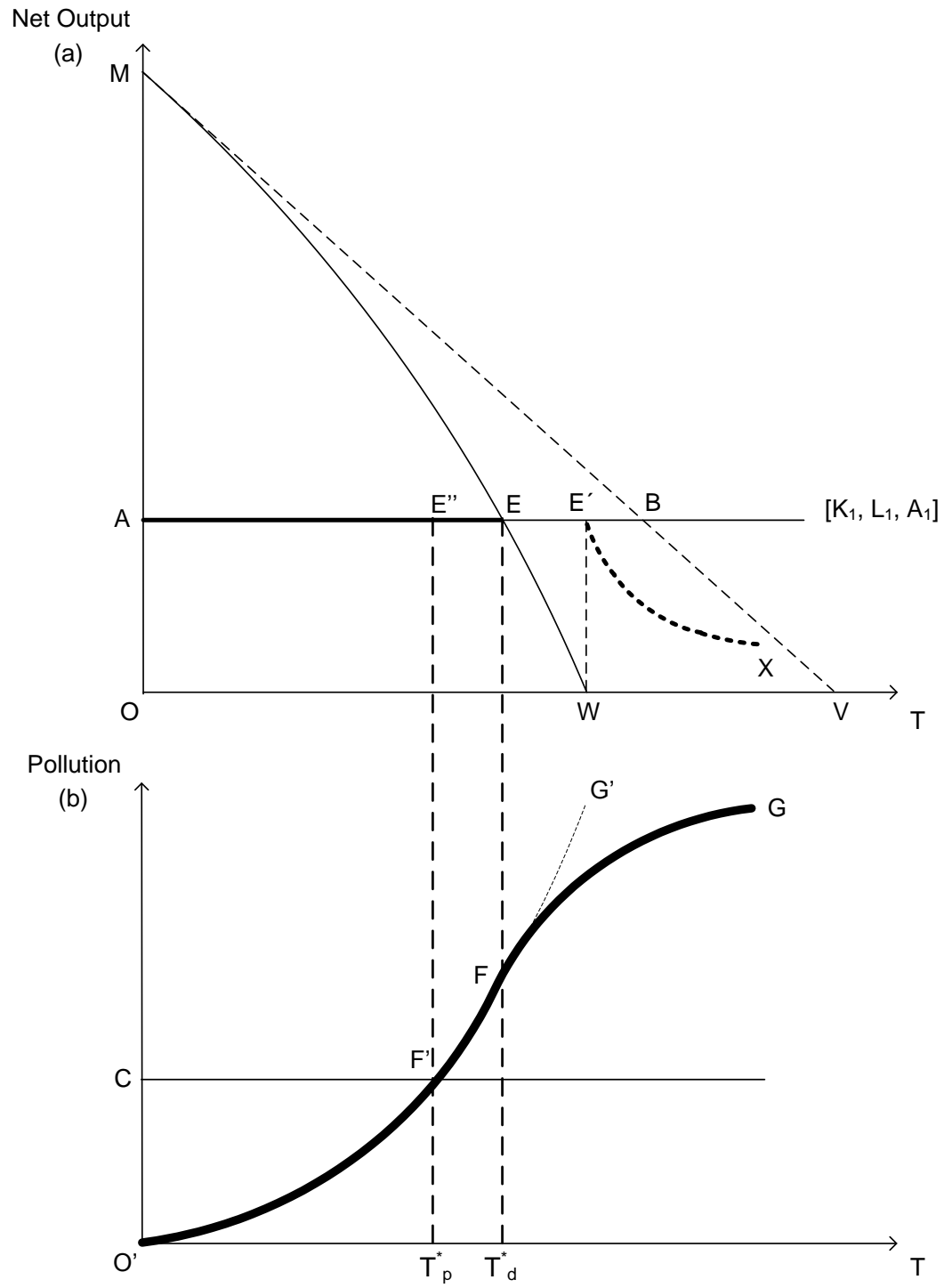


Figure 12.3. Limits to Growth under Evolutionary Dynamics.

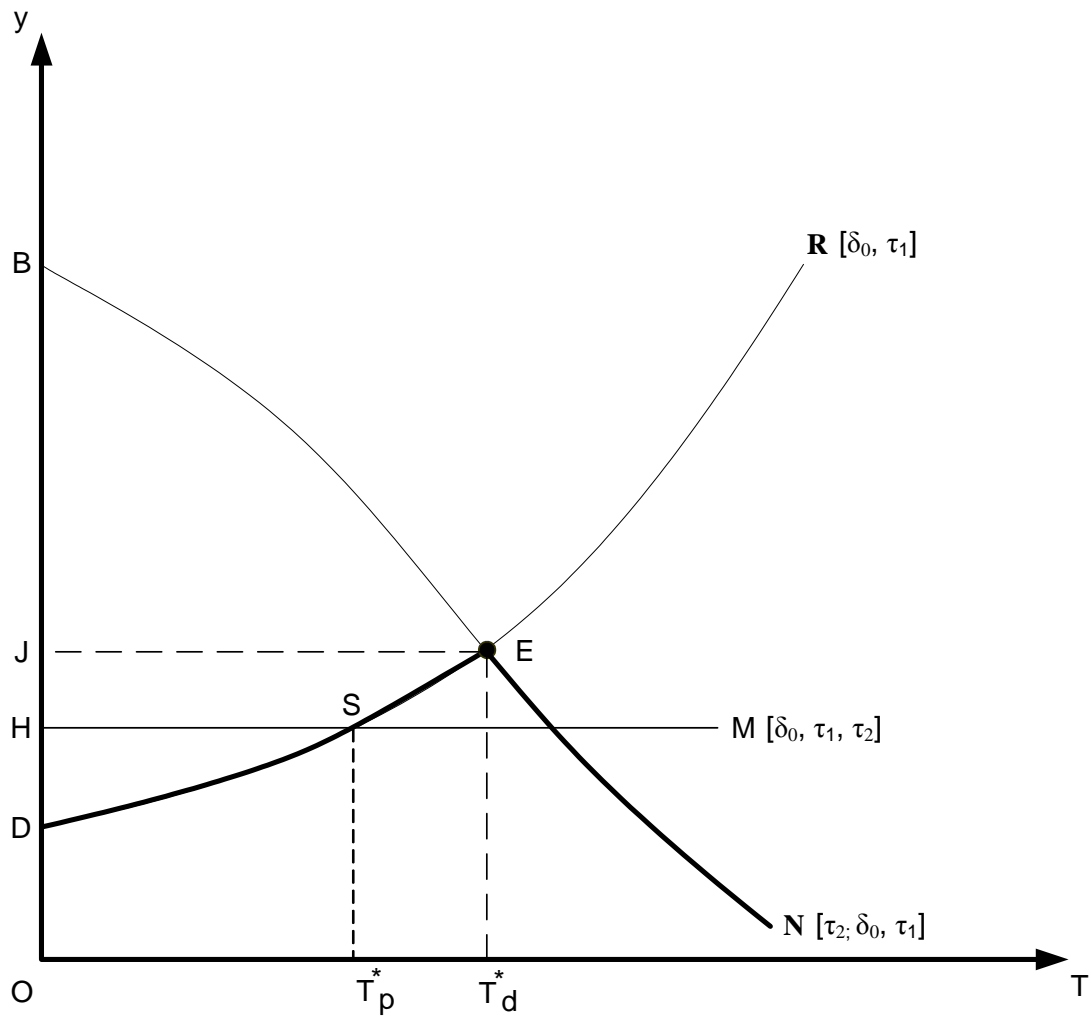
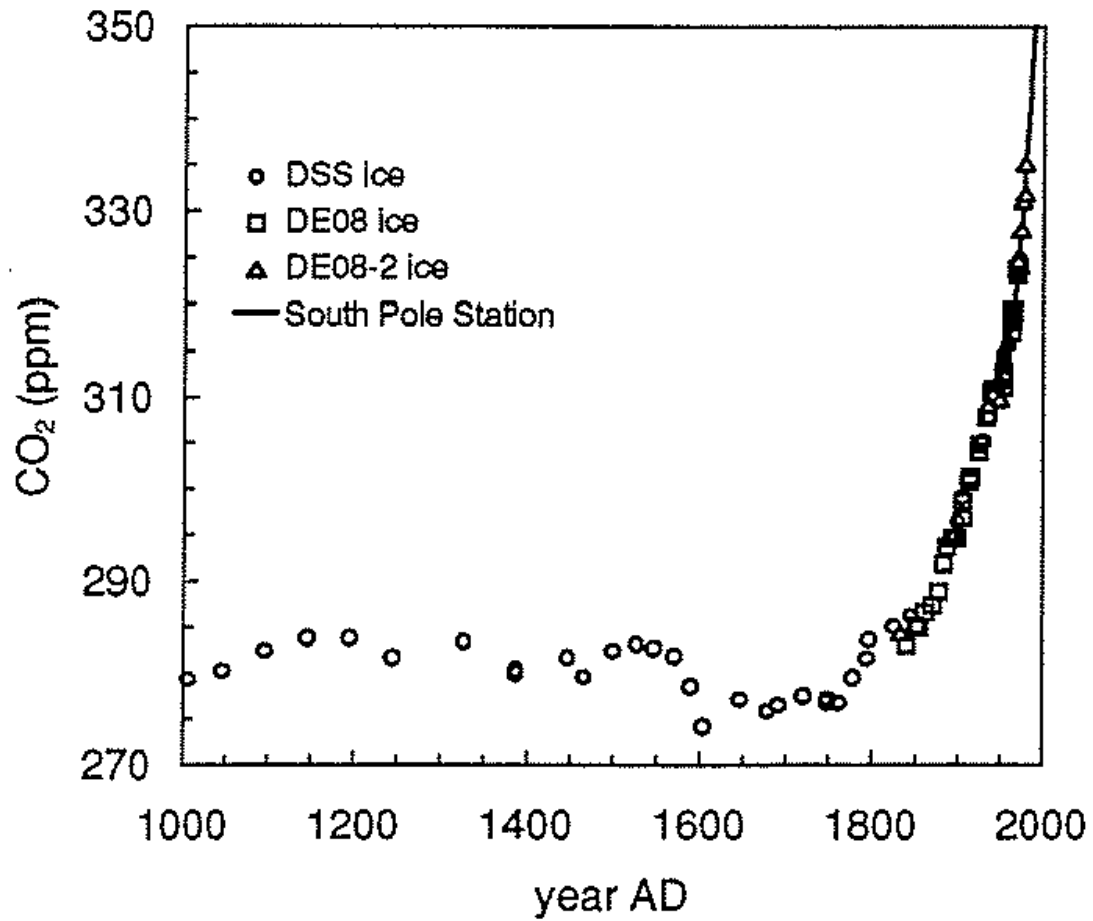
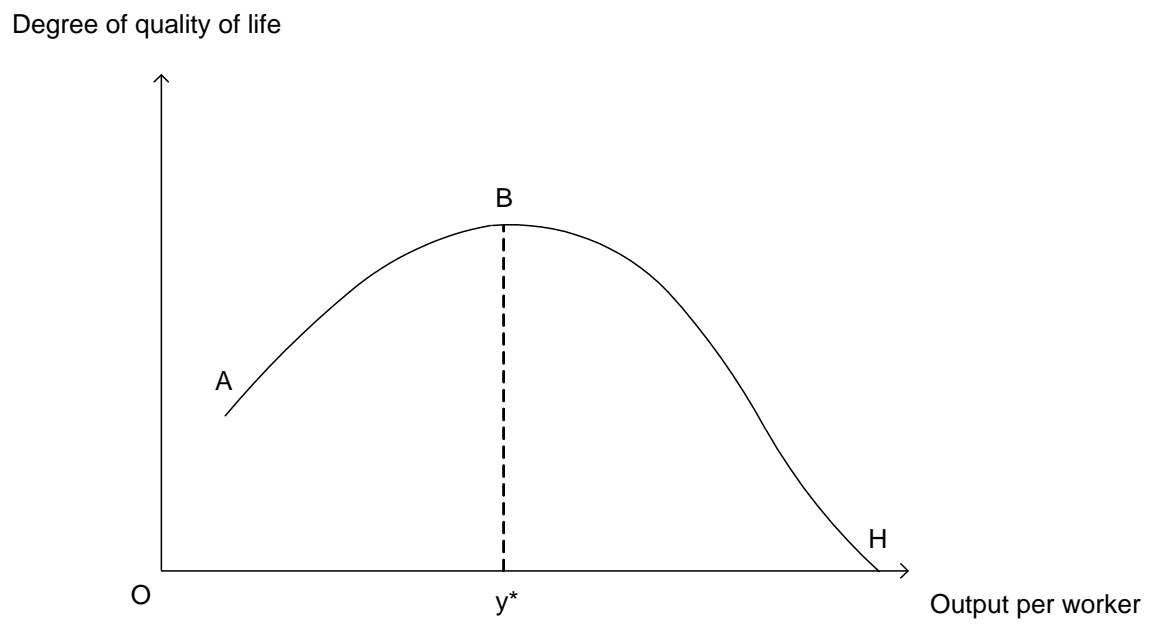


Figure 12.4. CO₂ concentration levels (part per million) from 1000AD to 1995AD



Source : Etheridge, et al (1996), p. 4123

Figure 12.5. Growth and Quality of Life

CHAPTER 13

A UNIFIED THEORY OF THE CAPITALIST SYSTEM

We have come a long way in the construction of a unified theory of the capitalist system. The rules of Popperian epistemology have been applied all the way. This chapter aims at summarizing the theory in the light of that epistemology. It will present its primary assumptions, predictions, and empirical falsification.

13.1 A New Economic Theory of Capitalism

The primary assumptions of the unified theory say that the capitalist system is constituted by different types of capitalist societies. Rich (First World) and poor (Third World) countries are not only quantitatively different, but they are also qualitatively different. Countries began to operate as capitalist societies under different conditions and at different times.

The unified theory assumes that individuals in a capitalist society participate in the economic process endowed with economic and political assets, which are unequally distributed among individuals. Because all types of capitalist societies are class societies, inequality in economic assets (land, physical capital, human capital) is common to all. Inequality in political assets means the existence of citizenship of different classes, with different entitlements in political assets, which implies a socially hierarchical society or heterogeneous society; when political assets entitlements are equally distributed, the society is said to be socially homogeneous. Then capitalist societies differ in their initial inequalities.

The second initial condition refers to their factor endowments. The unified theory assumes that capitalist countries differ in their initial factor endowments, the ratio of capital per worker. Capital includes physical capital and human capital. The value of the marginal productivity of total labor is the criterion followed to characterize capitalist countries. Technology is such that the higher the capital per worker endowment, the higher the value of the marginal productivity of labor will be. Then, if society's factor endowment is characterized by a sufficiently high capital per worker endowment, such that the value of the marginal productivity of the total labor force is positive, it is called under-populated; if it is characterized by a sufficiently low capital per worker endowment, such that the value of the marginal productivity of total labor force is zero or negative, it is called overpopulated.

Given those different initial conditions, the unified theory logically derives the proposition that capitalist countries have followed different economic development paths, in economic growth and income inequality. The ultimate factor that explains the existence and persistence of the high degree of global inequality in the capitalist system lies in the initial conditions; thus there is path dependence in the economic process of capitalism,

history counts. The unified theory is able to explain the whole set of known empirical regularities of the capitalist system. It is then a valid theory. This conclusion has been reached by following the scientific rules derived from Popperian epistemology, which was established in Chapter 1.

13.2 The Structure of the Unified Theory

The unified theory consists of three partial theories intended to explain production and distribution in the First World and the Third World countries. The latter is divided into two groups: those that have a weak colonial legacy and those with strong colonial legacy. The three theories are constructions of abstract capitalist societies characterized by two initial conditions only: factor endowments and initial inequality.

They have been named and characterized as follows:

- Epsilon: under-populated and socially homogeneous
- Omega: overpopulated and socially homogeneous
- Sigma : overpopulated and socially heterogeneous

These abstract societies intend to explain parts of the capitalist system, as follows:

- Epsilon society intends to explain the First World countries
- Omega society intends to explain the Third World countries with weak or no colonial legacy
- Sigma society intends to explain the Third World countries with strong colonial legacy, which constitute the large majority of countries in the Third World.

It has then been shown that a unified theory can be constructed from these three partial theories. Hence, a unified theory has been logically derived from the partial theories. The unified theory intends to explain production and distribution in the capitalist system, taken as a whole.

In order to make the partial theories and the unified theory falsifiable, particular models of these theories have been constructed. The structure of these models can be summarized as follows:

- Static models: Intended to explain the production and distribution process by types of capitalist societies and for the capitalist system as a whole; that is, to explain the observed difference in the *levels* of both income and degree of inequality.
- Dynamic models: Intended to explain the observed *persistence* of those differences in both levels in the process of economic growth.
- Evolutionary models: Intended to explain the *qualitative changes in the environment and in the human society* in the process of economic growth by introducing natural resources and the laws of thermodynamics in the economic process.

The eight empirical regularities listed at the beginning of the book (Chapter 2) are all consistent with the predictions of these models; that is, none of those models are refuted by the empirical regularities. The models indeed explain the realities that they intended to. At this stage of our research, therefore, there is no reason to reject the unified theory and it can be accepted provisionally, until new empirical regularities or superior theories appear.

13.3 Differences between Unified Theory and Standard Economics

By comparison to the unified theory, standard economics is unable to explain the eight basic empirical facts of the capitalist system. It can explain some of these facts, but not all. As indicated in the corresponding chapters above, standard economics can explain Facts 1, 4, 5 and 6, and Fact 7. But Facts 2 and 3 refute standard economics. In the short run, standard economics is able to explain unemployment in the First World, but not the coexistence of unemployment with underemployment in the Third World and in the capitalist system as a whole. Standard economics cannot explain Fact 8 either, as it predicts that economic growth has no limits.

Therefore, the unified theory outperforms empirically standard economics on explaining the given set of empirical regularities. The reason for this difference is to be found by using the scientific rules that were derived from Popperian epistemology in Chapter 1. According to these rules, empirical predictions are derived from the primary and auxiliary assumptions of a theory. Therefore, differences in predictions between theories must come from differences in their primary assumptions: differences in what these theories consider as the essential factors underlying the economic process. The set of primary assumptions of the unified theory is different from that of standard economics. The differences can be summarized as follows:

On the institutional context:

- The unified theory assumes the existence of three types of capitalist societies, *instead of* assuming just one, as is done in standard economics.
- Individuals participating in the economic process are endowed with economic assets (physical capital and human capital) and political assets, *instead of* assuming just economic asset endowments.
- The basic institutions of capitalism—markets and democracy—operate differently in each theory. The unified theory assumes a hierarchy of markets, in which some (e.g., labor markets) are more important than others for the reproduction of inequality, *instead of* assuming that qualitatively all markets play the same role in production and distribution. The unified theory assumes that democratic governments act guided by the motivation of self-interest of politicians; moreover, in sigma society, democracy operates with first and second class citizens, *instead of* assuming that government behavior is exogenous to the economic process.

On initial conditions

- Initial conditions include factor endowments and the initial inequality in asset distribution, *instead of* assuming factor endowments alone.
- The stock of non-renewable natural resources is given and production is subject to the laws of thermodynamics, *instead of* ignoring these constraints.

On economic rationality

- On the behavior of capitalists and workers, the unified theory assumes that they act guided by the motivation of self-interest, *as in* standard economics.
- Capitalists seek to maximize a hierarchically ordered preferences, in which the objective to remain as member of the capitalist class has the first priority over profit maximization, *instead of* assuming profits maximization as the unique objective.
- Political elites are assumed to act guided by the motivation of self-interest, as people do in any activity, *instead of* assuming that they act guided by the motivation of seeking social welfare maximization.
- Social tolerance to inequality is assumed to be limited, *instead of* assuming that it is unlimited.

This new set of assumptions constitutes the unified theory. It is, therefore, a new economic theory of capitalism. The three partial theories explain production and distribution in each type of capitalist societies, taken separately; whereas the unified theory explains the capitalist system, taken as a whole. Unity of knowledge, a requirement of scientific knowledge, is then achieved in the following sense:

- (a) In each partial theory, micro and macro behavior are integrated. General equilibrium analysis assures both the macroeconomic foundations of the micro behavior of social actors and the micro economic foundations of macro (aggregate) behavior of society.
- (b) In each partial theory, long run and short run equilibrium are integrated. Long run equilibrium is determined by the initial conditions of society; thus, long run equilibrium is determined independent of short run equilibrium, not vice versa. Therefore, short run equilibrium values fluctuate around the long run equilibrium.
- (c) The three partial theories are consistent with one another. The three economic theories can all be true. (This is opposite to the case of physics, in which quantum and relativity theories are inconsistent to each other, that is, they both cannot be true.) The corresponding unified theory can explain with a single unified theory a single ontological reality: the capitalist world.
- (d) The unified theory allows the construction of static, dynamic, and evolutionary models.

These characteristics are not found in standard economics. Macroeconomic models usually have no micro foundations, and microeconomic models rarely have macro foundations. There is ambiguity in the relationships between short run and long equilibrium, although most textbooks on macroeconomics also assume that the short run economic process does not affect the long run process (e.g., Blanchard 2009). The need of a unified theory is ignored in standard economics because it assumes just one type of capitalist society, in which there may be quantitative differences between countries, but qualitatively they are all similar, as if the real world were composed solely of omega and epsilon societies. One of the characteristics of the unified theory rests upon the introduction of the sigma society. Evolutionary models on the fate of the human species is absent in standard economics for it ignores the role of the entropy law of physics in the economic process.

In conclusion, the set of primary assumptions of the unified theory is able to generate a set of models from which empirical predictions have been derived. These predictions are consistent with the eight basic empirical facts about production and distribution in the capitalist system. The set of primary assumptions of neoclassical theory generate predictions that can explain some of these facts, but not all.

13.4 Explaining Overall Inequality in the Capitalist System

The unified theory is able to explain production and distribution in the capitalist system, taken separately by types of capitalist societies and taken as a whole. In the dynamic model, the analysis of production and distribution can be reduced to the distribution problem alone because differences in output per worker between the First World and the Third World can be seen as a distribution problem of total capitalist output between countries within the current generation. Because the unified theory introduces the laws of thermodynamics in the economic process, there is also the problem of inequality between generations, which was studied through an evolutionary model. Therefore, the unified theory can explain the overall problem of inequality in the capitalist system, which comprises three concepts of inequality: within-country, between-countries, and between-generations.

As to within-country inequality, the fact that Third World countries—with strong colonial legacy, which constitute the large majority—tend to be more unequal than First World countries is explained by their differences in the initial inequality, which is much higher in the Third World countries. This characteristic persists in the process of economic growth because growth cannot reduce endogenously the initial inequality, as shown in Chapter 11.

As to between-country inequality, the observed income gap between the First World and the Third World, and its persistence over time, is explained through the capital accumulation mechanism. Both groups of countries compete for private investment in international markets, in which Third World countries are in disadvantage due to their higher degree of social disorder, which originates in their higher degree of initial inequality. Relatively unequal countries are also in relative disadvantage to accumulate human capital and to retain it at home. As a result, the dynamic equilibrium differs between the two groups of countries. The position of the growth frontier curve of the First World is at higher level than that of the Third World, although with similar growth rates. Therefore,

convergence in income levels cannot happen endogenously in the process of economic growth, as shown in Chapter 11 as well.

The exception is the Third world countries with weak or no colonial legacy (few countries only), which will endogenously converge to the growth frontier of the First World. The reason lies in the nature of the initial inequality, which is qualitatively similar to that of the First World different, and thus differences are only quantitative in factor endowments. (This is the view that standard economics has of the capitalist system.) In the Third World countries with colonial legacy (large majority of countries), the difference with the First World is quantitative and qualitative: different factor endowments and socially heterogeneous.

The unified theory can therefore explain both the within-country and between-country inequality (Facts 6 and 7, listed in Chapter 2). The initial inequality in the distribution of economic and political assets within the capitalist system is the ultimate factor that explains the long run global income inequality in the capitalist system.

According to the unified theory, in the process of economic development of nations, there is path dependence; that is, history counts. The existence of path dependence implies that in society the past is not past, but it is still with us in the present; that is, in some ways, “society has no history.” The persistence of some colonial institutions in the capitalist system of today, such as classes of citizens, is a good example. The workings of the capitalist system cannot reduce inequality endogenously; there are no mechanisms in the capitalist system that can make capitalism produce a less unequal society. The process of economic growth cannot endogenously break with history. When path dependence exists in the economic process, history is not a curiosity, but a necessity to understand the present. Economics then becomes a truly historical social science.

The conclusion that the economic process shows path dependence does not imply historical determinism, however. The economic process is not like that of physics in which the laws of the universe originated from the big bang and nothing can be done about its trajectory now. In economics, there are exogenous variables, which could be used in public policies to break with history. This will be shown below.

As to intergenerational inequality, we must use the evolutionary model. This model assumes that the laws of thermodynamics are essential in the economic process. It also assumes a single world society. The output of the economic process includes a good and a bad; the good takes the form of private consumption good and the bad is a public bad: waste and pollution.

The evolutionary model predicts that as economic growth proceeds, qualitative changes will occur in the environment, namely, continuous and irrevocable degradation. Economic growth cannot continue forever. The evolutionary model is able to explain the fact that economic growth implies continuous degradation of the environment (Fact 8). The consequence is that the human society faces new economic problems, such as the limits to economic growth, the intergenerational inequality, and the quality of life.

The predictions of the evolutionary model include limits to economic growth and therefore an increase in the intergenerational inequality. As more non-renewable resources are depleted in our generation, less quantity of resources are left for the future generations,

who will not be able to replicate the consumption level of our generation. Quality of life will not be any better in future generations than it is in our generation either.

Regarding overall inequality—within-country, between-countries, and inter-generations—and quality of life in this generation and next generations, the unified theory shows that economic growth is neither a sufficient condition nor a necessary condition to achieve social progress. It is not sufficient because economic growth has not produced social progress; it is not necessary because social progress can be achieved without economic growth. The pronounced inequality at the world scale implies that improvement in the quality of life of the masses through redistribution is an alternative to economic growth. This result has consequences for public policies, as will be discussed in the next chapter.

In sum, the empirical predictions derived from models of each partial theory and from the unified theory are consistent with the eight empirical regularities found in the workings of the capitalist system (Chapter 2). Therefore, the partial theories explain production and distribution in the First World and Third World countries, taken separately, and the unified theory explains the capitalist system taken as a whole. Thus, there is no reason to reject the unified theory, and we can accept it at this stage of our research, until new empirical regularities or a new superior theory appears. The unified theory can then be taken as the basis to discuss public choices and to derive science-based public policies.

13.5 Comparisons with Other Theories

Some differences between the unified theory and other economic theories of the capitalist system can also be summarized now. Since the unified theory deals with two categories of inequality that are observable, the comparison will be carried out for each category.

Within-country Inequality

As to within-country inequality, an important common trait of capitalist societies is, according to the unified theory, the use of labor discipline devices to extract effort from workers. Unemployment constitutes such device in the First World and underemployment in the Third World. Both devices create inequality among those workers that are endowed with similar human capital. Therefore, in the unified theory, inequality among workers constitutes the generalized labor discipline device. Inequality is thus embedded in the functioning of labor markets. This is the unified theory of the labor market in the capitalist system. By contrast, the standard labor market theory tends to concentrate on the unemployment problem alone in all types of capitalist societies.

The results of the unified theory are in agreement with some predictions of the modern theories of labor exploitation. The labor market and the credit market constitute the two mechanisms under which workers can be exploited, as Roemer (1982) has shown. Therefore, workers who are excluded from these markets are not exploited. Joan Robinson's well-known dictum "There is only one thing that is worse than being exploited by capitalists. And that is not being exploited by capitalists" is in accord with the conclusions of the unified theory, if all forms of exclusion, not only unemployment, are

considered situations of not being exploited. According to the unified theory, capitalist societies with higher degrees of inclusion (higher degrees of exploitation) show lower degree of inequality than those with lower degree of inclusion (lower degree of exploitation).

Standard textbooks tend to justify inequality as part of the incentive system needed in the economic process. “The rich deserve the income they get.” Profits are the rewards to creativity, innovation, and entrepreneurship, and risk taking. Similarly, high salaries are the rewards to high investment in human capital. The unified theory shows that income inequality is caused by the initial inequality; hence, the lower the initial inequality, the lower the degree of income inequality will be. Profits and high salaries would still be the rewards to the factors indicated above, but income inequality would be lower because capital assets (in the form of physical capital and human capital) would be less concentrated; that is, the same rewards would reach more people.

Some view inequality as unrelated to the incentive system: “The rich do not deserve the income they get.” The argument is that profits mostly come from rents; that is, from monopoly, capture of the state, rents from social networks, and even illegal practices and corruption. According to the unified theory, rents are endogenous; they are derived from the initial inequality, from capital concentration.

Between-country Inequality: The Role of International Trade

As to between-country inequality, according to the poverty trap, or low-income trap theory, the disadvantage of the Third World is their low-income, from which savings can be but small. According to the unified theory, the trap lies not in their poverty level, but in their initial degree of inequality in the distribution of economic and political assets, which is a crucial factor in the determination of investment.

Neoclassical international trade theory assumes that the comparative advantage of countries originate from differences in factor endowments. The theory predicts that trade of goods will equalize factor prices across countries. Moreover, equalization of real wage rates will result from trade. This prediction has been refuted by facts.

In contrast, the Ricardian trade model seeks to explain trade by differences in labor productivity, which is the source of comparative advantage. The model assumes that labor is the only scarce factor of production and that the quantity of labor needed per unit of output is fixed (exogenously determined). The model predicts that differences in real wage rates after trade will tend to be proportional to the differences in labor productivities. Facts are consistent with this model, as reported in popular textbooks on international economics. The Ricardian model is however too limited as to the role of inequality in international trade.

As shown in Chapter 10, the trade model derived from the unified theory is consistent with the generalized Ricardian model, in which labor is not the only scarce factor of production and average productivity of labor is exogenously fixed but in the sense of the *level* of labor productivity (a given labor productivity curve), along which short run variations can occur. The unified theory predicts that, after trade, countries differ in their levels of labor productivity, which also implies differences in the level of real wage rates:

countries with high levels of labor productivity also have higher real wage rates. There cannot be real wage rate equalization across countries.

Now countries with low-wages and low-productivity can compete in international markets. Suppose that the First World has absolute advantage in both goods, B and C, but comparative advantage in good B; that is, its relative labor productivity in B is higher than it is in the Third World. The First World may have a cost advantage in good B, despite its higher wage rate, because the higher wage is more than offset by its higher labor productivity. Similarly, because of its lower wage rate, the Third World can have a cost advantage in good C, even though it has lower labor productivity. What is relevant in the working of the comparative advantage mechanism is the cost advantage.

What factors do determine the level of labor productivity or output per worker across countries? Chapter 11 answered this question. These factors are technology, factor endowments, and initial inequality. In the entropic economic process of Chapter 12, it can be shown that the factors are the same. In the long run, changes in these factors depend upon the investment in physical capital and human capital (and in infrastructure as well), which depend upon the rate of investment of each society. But private investment and public investment across countries depend upon social order of societies, which in turn depend upon the degree of inequality in the distribution of economic and political assets.

Capitalists decide in which country and industry to invest. They may want to exploit the natural resource endowment of countries or their human capital endowment or their external economies. Investment in one industry will increase labor productivity in that industry, as workers will be equipped with more capital and higher technology. Hence the international trade mechanism would operate as follows: private investment increases labor productivity in the industry, which in turn creates or reinforces its static comparative advantage, and which attracts more investment, and so on. This dynamic model now says: *It is not that comparative advantage brings in private investment to the country; on the contrary, private investment develops comparative advantage by increasing labor productivity.*

The question is then, what are the determinants of the allocation of private investment across countries? According to the unified theory (Chapter 8), the essential factor that determines the allocation of private investment across countries is the degree of inequality of countries. More unequal countries are riskier countries for private investment because higher degree of inequality generates higher degree of social disorder. Therefore, the unified theory implies the following international trade mechanism. Firstly, countries compete in the international arena for private investment, and in this arena compete with their degree of social order and the underlying degree of equality. *It is not that trade determines income distribution, as neoclassical theory says; on the contrary, it is differences in the degree of inequality of countries that determines international competitiveness and trade patterns.*

Secondly, in the explanation of the observed patterns of international trade, differences in relative labor productivity constitute the *proximate* factor only, because in the long run these differences are endogenous, as they depend upon investment on physical capital, human capital, and higher technology with which workers are equipped in each society and industry. The *ultimate* factor is the difference in inequality between countries, which determines the allocation of investment across countries and industries, as the

unified theory predicts. Consequently, the unified theory predicts that international trade of goods cannot equalize the level of real wage rate (for each level of skill)—for that would require equalization in the level of labor productivity—between the First World and the Third World, as long as the differences in the degree of inequality remain unchanged.

On the other hand, neo-marxian models argue that trade implies the exploitation of the Third World by the First World through international trade and the price mechanism (through international terms of trade or wage differentials). This model predicts that exploitation between countries would have to operate through profit remittances of foreign direct investments; however this effect is small. Indeed, profits generated in the Third World and remitted to the First World represents only 0.5% of the latter's GDP and 1.5% of the Third World's GDP (IMF 2011, Tables B13 and B16). The explanation given by the unified theory to this small effect is that foreign direct investment originates mostly in the First World and is allocated mostly to the same First World (Chapter 8), which implies that profit remittances from foreign direct investment move mostly within the First World countries.

Between-country Inequality: The Role of Initial Conditions

Initial conditions play an essential role in the unified theory. Other theories of initial conditions run strong in the literature. One is geography. The persistent gap in the income level between the First World and the Third World can be explained by geographical initial conditions. Western Europe superiority can be explained by geography: the natural resource endowments, climate, and accessibility. For this reason, it was the European who conquered the inhabitants of the New World, and not vice versa (Diamond 1997).

Some research on economic growth takes into account the initial inequality in the individual distribution of assets. Galor & Zeira (1993), Galor & Moav (2004), and Galor (2011) develop models in which more equal initial inequality in the distribution of physical capital is good for growth. The mechanism operates through human capital accumulation. The empirical falsification of these models is pending, but the authors argue that available empirical studies tend to corroborate the empirical predictions, mostly for the First World. By comparison, the unified theory assumes the initial inequality as composed of three assets: physical capital, human capital, and degree of citizenship; moreover, the empirical predictions refer to differences between the First World and the Third World and show consistency.

The other initial conditions refer to institutions. Some researchers argue that initial institutions explain the income level inequality. In the colonial period, mortality rates of European colonizers were very high in some territories, but low in others. In the former case, the economy was mostly extractive of local resources while in the latter property rights were established and these are the high income countries of today (Acemoglu et al 2001). Others consider that human capital endowment is the factor that develops institutions that are favorable to democracy and property rights (Glaeser et al 2004).

A more recent study by Oded Galor (2011) argues that income level differences can be explained as follows: rich countries reached the take-off stage of growth *earlier* than poor countries, which in turn is explained, at least in part, by institutions; hence, the

implication of the model is that between-countries inequality is just temporary and absolute convergence should take place in the future. This remains to be seen.

The unified theory proposes the degree of inequality in the distribution of economic and political assets among individuals as the essential initial condition. This initial inequality is what generates path dependence in the process of economic growth. Geography plays a role in this explanation. The initial inequality has to do with the colonial history of countries, which refers to European colonialism. But European superiority had to do with geographical factors, as Jared Diamond argues. Geography indeed plays a role in the unified theory, but through colonialism.

Institutions are endogenous in the unified theory; hence institutions cannot explain capitalist behavior. The development of democracy and property rights in society depends upon the initial inequality. In very unequal societies, the economic process produces social disorder, which set limits to the development of these institutions. The informal economy is as important as the formal economy (operating under the rules of property rights and democracy); hence, illegal activities, corruption, political instability are all endogenous. The outcomes of the education process, human capital and citizenship, are also endogenous in the unified theory.

A Summary

Confrontation of the *assumptions* of different economic theories with each other cannot produce scientific knowledge. The scientific rules indicate that knowledge emerges from the confrontation of the *empirical predictions* of the theories with the available empirical data, as shown in Chapter 1. According to this epistemological criterion, the predictions of the unified theory are consistent with available facts about production and distribution in the capitalist system, which includes inequality within-country and between-countries. So the theory may be accepted. The competing theories do not constitute a unified theory—seeking unity of knowledge—of production and distribution of the capitalist system, and as such have not been submitted to the falsification process either.

Changes in the economic performance of the Second World, particularly China's, in the last decades has been ignored in the unified theory. Those changes have been implicitly assumed as exogenous variables and with marginal effects upon growth, inequality, and environmental distress in the long run period of analysis (since the 1950s). The fact that the empirical predictions of the unified theory are consistent with facts indicates that this assumption is roughly correct. However, a “grand unified theory” would be needed to explain the relations between growth, inequality, and the environment in the more recent world economy, in which the interactions between the First World, the Second World, and the Third World would be the subject of theoretical analysis and empirical confrontation.

13.6 The Capitalist System as Sigma Society

What type of society is the capitalist system, taken as a whole? Could it be seen as a sigma society? If the answer is yes, we will have an analytical advantage because we already know how a sigma society functions, what the exogenous variables are, and then we could

know what the instruments for public policies are. This section then seeks to show, as a first step to discuss public policies, that the capitalist system as a whole can indeed be seen as a sigma society.

The dynamic model of the unified theory assumes capital mobility across societies, but not labor mobility (Chapter 11). However, internal and international migrations are facts of life. The model has thus ignored these facts. The current international migration from the Third World to the First World seems to follow a period of rapid internal migration (rural-urban) within Third World countries. Income level disparities are the possible factor explaining both movements.

Surely, models of the unified theory that include factor labor migration as an essential factor could be constructed. Fact 6 and Fact 7, however, indicate that internal migration does not seem to shift in any significant way the degree of inequality in the capitalist system. Internal migration within the Third World does not seem to have significant effects in reducing within-country inequalities. These same Facts indicate that, similarly, international migration does not seem to have significant effects in reducing the income inequality between countries. The income level disparities are persistent.

In both cases, the reason would be that migration (internal or external) does not reduce significantly the inequality in asset distribution, which is the essential factor explaining income inequality. The voice of the poor when living in the city may be louder than when living in the countryside, but its effect in reducing citizenship inequality does not seem to be significant.

The model of the unified theory that has been presented above was able to predict Facts 6 and 7. If it had not, we could have concluded that some essential factors were left out in the model, such as migration, and that new models taking migration into account were needed. To be sure, the unified model ignores the effect of labor migration upon inequality, but it does not mean that such effect does not exist; it only means that the theory assumes that the effect does exist, but it is not significant, and can thus be ignored. Facts do not contradict the theory.

We know that a significant proportion of workers migrating from the Third World to the First World have the status of illegal workers; that is, they are second class citizens in the First World. This migration process may then be transforming the First World into a sigma society. In the long run, the world capitalist system may then become composed of only sigma societies. The unified theory of capitalism would be just the sigma theory.

Even today, the inequality in asset distribution within and between countries implies a very unequal capitalist society in the aggregate. The degree of concentration of physical capital seen at the world capitalism scale is very high. The same can be said about human capital. In terms of political assets, again, the degree of inequality is also very pronounced. Most workers of the Third World can be seen as second-class citizens at the world scale, which implies that the z-workers of the Third World could then be seen as third-class citizens at the world scale. The capitalist system is socially hierarchical and overpopulated and therefore corresponds to the concept of a sigma society.

The capitalist system may be viewed as operating with the rules of capitalism but together with some rules that are legacies of the colonial system. Private property, markets, and democracy, together with hierarchy of citizenship classes, constitute the institutions of

the capitalist system. The process of economic growth has not been able to break with history. These are characteristics of a sigma society.

Inequality is not only about quantitative differences between social groups; it is also about qualitative differences. In each type of capitalist society, the rich not only has more money than the poor, but they live a life that is qualitatively different. Income differences imply qualitative differences in consumption baskets, neighborhoods, and cultures. Income groups are different social groups. Between capitalist societies, First World countries are not only richer than Third World countries, but they live a life style that is qualitatively different. The capitalist system is thus composed of “worlds apart” at both national and international levels, just as a sigma society.

On the empirical significance of seeing the capitalist world as sigma society, British historian Eric Hobsbawm (2002) makes the observation that nearly 80 percent of people in the Third World live outside the zone inhabited by people with white skin. “These 80 percent knew nothing of the world and, give or take a few thousand individuals, the world knew practically nothing about them” (p. 363).

The unified theory predicts that income inequality depends upon the initial asset inequality. If the latter is higher in the capitalist system than in either the First World or Third World taken separately, so will income inequality. As we know from statistics, the variance of two distributions is not equal but higher than the weighted sum of the variances of each distribution. Similarly, we can say that the degree of inequality (measured by the variance or other index) of two distributions is higher than the weighted sum of the degree of inequality of each distribution; hence, the aggregate inequality of two distribution could be higher than each inequality, but never equal or smaller than each.

The Gini index for the inequality in the distribution of the global income *among individuals of the world* (regardless of the country of residence) is estimated around 0.64 for the period 1988 to 1998 (Milanovic 2005, Table 9.4, p. 108). This measure includes China, which is not considered part of the capitalist system in this book. Adjusting by the China effect (Table 9.6), the Gini coefficient would be reduced to near 0.55. By comparison, the average Gini index for the First World was 0.33 and 0.47 for the Third World in the period 1950 to 2008 (Table 2.2). Indeed, the observed global income inequality in the capitalist system is higher than the observed inequality in the First World and also higher than that in the Third World.

Political scientist Samuel Huntington has developed a theory on the role of culture and religion in the political and economic development of societies. A prediction of his theory is that religious conflicts now (after the collapse of the communist world and the end of the cold war) constitute a major factor for world peace (Huntington 1996, p. 321).

The unified theory suggests that there is another source for violence in the world: it is the persistence of a pronounced inequality at the world capitalist scale. Inequality plays a role in shaping the quality of society. We live in an increasingly globalized world, especially regarding communications. The income disparities at the world scale are then more flagrant than ever. The higher the degree of inequality in society, the higher the social disorder will be. The total social disorder observed at world scale can be seen as the result of two effects: part is due to the within-country inequality and the other part to between-country inequality, between the First World and the Third World.

A summary of the unified theory is in order. The basic causality relation established by the unified theory is that the three outcomes of the economic process—growth, inequality, and environment degradation—are explained by the initial inequality under which capitalist countries were born, which could be called the *power structure of society*. According to the unified theory, most market prices and quantities are endogenous variables; government policies are also endogenous; technological progress is also endogenous, because all these variables are determined by the initial inequality. Therefore, the exogenous variable explaining the economic process is the initial inequality of countries and the initial inequality of the capitalist system as a whole. Moreover, there is no mechanism in the economic growth process than can reduce endogenously the initial inequality, the power structure of society. The initial inequality is possibly not the only ultimate factor; but according to the unified theory, it is the essential factor. These causality relations—a set of beta propositions—has been proven to be consistent with the empirical regularities of capitalist development.

How could the power structure of society get changed or how to break with history? This is the main public policy question that comes from the unified theory. The analytics of this question—science-based policies—will be presented in the next chapter.

CHAPTER 14

SCIENCE-BASED PUBLIC POLICIES

Science-based public policies need positive and normative sciences. Positive science is what this study has presented so far. It is another name for factual sciences. Normative science is a formal science and studies the logical system about ethics, values, and norms. In short, positive science seeks to answer the question *how the world is*, whereas normative science seeks to answer *how the world ought to be*. Science-based policies are based on scientific knowledge about the social world, but not on this scientific knowledge alone, for normative alternatives will also emerge from the policy implications of scientific knowledge.

Positive and normative economics are different sciences and both are needed to discuss science-based public policies. The basic reason is that there is no one-to-one relation between the causality relations established by a valid economic theory and the policy implications of the theory. Only in the case of a theory with one endogenous variable and one exogenous variable such relation would be one-to-one; for there would be just one objective to seek and just one instrument to use. In the more general case, of several endogenous variables and several exogenous variables, alternative objectives and alternative instruments are logically derived from the theory. In this general case, there is no logical route from the proposition *how the world is* to the proposition *how the world ought to be*. Normative propositions are needed in between. This principle will be seen in action in this chapter.

14.1 Growth versus Quality of Life: Today's Big Trade Off

The unified theory has been able to explain production and distribution in the world capitalism, taken separately and taken as a whole. World capitalism resembles a sigma society, as was shown above. Science-based policies will then be discussed taking into account the relationships discovered by the unified theory.

Several basic causality relations in the long run functioning of the capitalist system have been discovered in this book. First, the process of economic growth has been accompanied by the increase in the global income inequality, which includes within-country, between-country, and intergenerational inequalities. Although within-country inequality has remained nearly stable, between-country inequality has increased, and intergenerational inequality will. Second, economic growth has been accompanied by increasing shocks upon the economic process, which tend to reduce quality of life; the shocks refer to social disorder (due to high degrees of within-country inequality and between-country inequality) and to pollution and climate change. Third, economic growth has been accompanied by increasing pollution and increasing rate of depletion of the non-renewable resources, which endanger the survival of human species, as we know it.

In sum, in the process of capitalist economic growth, overall technological progress and modernization, together with increasing quantities of material goods produced, have taken place; consumption per capita has increased over time and has contributed to a higher quality of life. However, economic growth has side effects which tend to affect negatively the other components of quality of life: social disorder, inequality in consumption, higher risk of environmental death and morbidity. The net effect of economic growth may be positive now, but will tend to be negative as growth proceeds. Hence social objectives are in conflict and social choice is necessary. Growth versus quality of life constitutes the big trade off in our time.

What are the policy instruments? Changes in the exogenous variables of the unified theory, which includes institutions, technological change, and the initial inequality, will change the endogenous variables growth and quality of life. The public policy implications of the unified theory will be discussed in this final chapter.

14.2 Market or Democracy Failures?

Which of the fundamental institutions of capitalism—market or democracy—are responsible for the social progress failure in the process of economic growth?

The famous statement by Adam Smith (1776)—the behavior of everyone acting guided by the motivation of self-interest will lead, as by an invisible hand, to the common good—constitutes the justification of market institutions. Leon Walras (1883) and other mathematical economists showed later on that this statement was a theorem; that is, it was valid under certain conditions only. The conditions include perfect competition, absence of externalities, and perfect information; most importantly, markets must be Walrasian.

Under these conditions, market prices will reflect scarcity of resources, that is, market prices will reflect the marginal *social cost* of producing goods (not the marginal cost to the firm), which is a criterion of economic efficiency. If producing one additional unit of wheat implies giving up two units of corn in society, then the relative market price of wheat to corn should be double and the market system will indeed generate this relative market price. Thus the production outcome of the market system will be economically efficient.

If one of those conditions is not met, individual self-interest cannot lead in the aggregate to the efficient utilization of scarce resources, to economic efficiency, to the common good; then it is said that the market system fails. As we have shown above, some important markets—called here basic markets, including labor markets—are not Walrasian. It is also a fact that the real world market system does not operate under perfect competition, due to the generalized existence of monopolies and oligopolies, in which case the market price of a good will not be equal to its marginal social cost of production.

There are also significant cases of externalities, particularly in regard of exploitation of mineral resources, in which the costs inflicted to other sectors (such as agriculture) are not internalized in the production cost and market prices; hence, market prices cannot reflect the marginal social cost of producing minerals. Market prices do not reflect the damage on the health of people due to pollution either. Finally, market prices do not reflect the preferences of all the social actors involved in the allocation of economic resources,

which is another criterion of common good, because the preferences of the future generations in the allocation of non-renewable natural resources are absent.

Another assumption of the market efficiency theorem refers to the role (or no role) of income inequality. Income inequality is not part of the criterion of the common good. Any degree of income inequality that results from market exchange is acceptable as efficient. Income inequality is not part of market failure. To be sure: Pareto optimum can imply a high degree of inequality.

The logic of the market general equilibrium theorem is simple. Given the initial asset endowments of individuals, given the private property of these assets, and given the norms of market exchange, and given perfect competition, absence of externalities, and perfect information, prices and quantities of equilibrium will be determined, resource allocation will also be determined, which will be efficiently allocated. Income inequality will also be the outcome of the market exchange, but it is not considered part of the market outcome evaluation.

The market system will be considered efficient, independently of the outcome over the degree of income inequality. Moreover, if income inequality is too concentrated in few hands, the market system has no mechanism to reduce endogenously this high degree of inequality. In contrast, if goods in Walrasian markets are in excess demand or excess supply, the market mechanism will endogenously eliminate those excesses; thus Walrasian markets are self-regulated. The role of the market system is to solve for prices and quantities of equilibrium; and the solution is, under certain conditions, economically efficient. But income inequality is not self-regulated because it is the outcome of market equilibrium of prices and quantities. A high degree of inequality that is beyond the socially tolerable inequality may still be economically efficient; however, it is not socially efficient, as it generates social disorder and lower quality of life for all; yet it does not constitute market failure.

Is the market system to be blamed for what is called “market failure”? As shown along this book, the task of markets is to solve for prices and quantities of equilibrium, of general equilibrium, which includes Walrasian, quasi-Walrasian, and non-Walrasian markets. The economic theory of markets assumes that the market system operates *as if* it solved a system of equations; the market system operates like an equations-solving machine, just like a big computer.

The market system will then solve the equations in which social interactions are represented. If the solution is not economically efficient or is not socially desirable, we cannot say that the market system has failed because the equations-solving machine has done its job: it has found the solution to the system of equations and has produced the prices and quantities of equilibrium. Market failure would occur if for some reason the system of market exchange could not find the prices and quantities of equilibrium; that is, it would like a broken computer.

What is called “market failure”—economic inefficiency or socially undesirable solutions—does not have to do with the computer. It has to do with the equations to be solved, which reflect the underlying social interactions and distributional social conflicts in society. Social interactions and distributional conflicts operate through market exchange. It is not that the market system determines the power structure of social interactions.

If individuals interact in the market under the particular context in which asset endowments are unequally distributed among people, which implies economic and political structures that are characterized by high degree of power, then there will be a particular market solution of prices and quantities and income distribution; if power structure were less concentrated, then there will also be market solution of prices and quantities, but they will be a different solution: other prices and quantities and thus other income distribution. If democracy intervenes and sets additional conditions, the market system will still come up with prices and quantities of equilibrium and a degree of income distribution; if democracy does not intervene, the market solution will be different. The market system cannot be held responsible for these different solutions.

Free-market in the ordinary language simply means no state intervention. But then monopoly and oligopolistic market structures, which reflect high concentration of physical capital in few hands, would also be free-markets. If the market solution of this “free-market” structure is inefficient or socially undesirable, the problem is not the market system, but the market structure, the power structure, which determine the conditions under which the market system will operate and solve for prices, quantities, and income inequality. An analytical distinction needs to be made between *market system* (the equations-solving machine, the computer) and the *market structure* (the degree of market power and the corresponding equations to be solved). The market system is the servant, not the master; the market structure is the master.

But there is confusion between these categories of market system and market structure. Very often we hear the economic elites saying that “the market is worried” about some interventions, indicating that they see the market system as part of their property rights. We also hear very often workers also complaining about the market system. Certainly, the enemy of workers cannot be the market system; it is the power structure underlying the market system. A society of workers only would still use the market system. The market system delivers a degree of income inequality as *output* depending on what went in as *inputs*.

Consider a capitalist society in which the initial inequality is very low. In such a society, the market system would deliver a very low degree of income inequality; that is, this is the same market system that delivers the high degree of inequality that we observe today. The market system is socially blind and short sighted. Actually, as noted above, the market system can be seen as servant, not as the master; hence, market failure originated in the market structure, in the high concentration of physical capital.

Democracy is the other basic institution of capitalism. Democracy (from the Greek *demos*, people and *kratos*, government) is a form of government in which people have an equal saying in the decisions that affect their lives. It is attributed to American President Abraham Lincoln the best definition of democracy: It is the government of the people, by the people, and for the people.

While the market system is the mechanism used for the solution of the problem of production and distribution of private goods, democracy is intended to be the mechanism to solve the production and distribution of public goods. What are the factors that determined the quantity of public goods to be produced? What goods will be taken out of the market and treated as public goods? How are they going to be produced and distributed? How to intervene in the market system?

The answers to these problems are reached through social choice and democracy is the mechanism for that. Democracy is explicitly the mechanism for the common good. However, democracy is a political system that can deliver different outputs, depending upon who controls the state. Voters do not control the state. Usually voters do not decide on social issues directly, but only indirectly, through the elections of their representatives, which is a way to generate political power. Democratic governments are run by politicians, who also act guided by the motivation of self-interest (Chapter 7). Politicians are short sighted, as they seek to maximize votes for the next elections. Thus the principal-agent problem appears: the agent (government) has no incentives to do what is convenient for the principal (voters). In addition, voters have the incentive to act as free-riders on public matters. "Everybody business is nobody business." Thus the equilibrium of the democratic system is also reached: politicians can act in this way period after period.

But, again, the outcome of the democratic mechanism will depend upon the power structure underlying the democratic process. For example, and as implied by the unified theory, a democratic mechanism in which there are first class and second class citizens will turn up a different solution compared to the case in which voters are all first class citizens.

As in the case of the market system, failure of democracy has nothing to do with the democratic system itself. The democratic mechanism will process what the dominant power structure will feed in the system; if the power structure were changed, the outcome will be different. Markets and democracy will process what the balance of power between capitalists, workers, and governments feed in in the production and distribution of private and public goods. Failures originate in the nature of social interactions, in the power structure underlying the social interactions, which in turn come from the *inequality* in the distribution of economic and political assets.

In sum, the dynamic equilibrium with rapid economic growth, together with persistent high degree of inequality and increasing risk in the life of people, which combined affect negatively the quality of life, that has characterized the capital system, is caused by the tremendous power concentration in society, a reflection of the high degree of inequality in the distribution of economic and political assets, both at national and international levels. This is an exogenous variable of the unified theory. The socially inefficient solution is due to the current power structure, not to market and democratic institutions.

Could we say that the socially inefficient solution is a failure of capitalism? No. Capitalism is a class society. Capitalism is a society in which people act guided by the motivation of self-interest. The objective of capitalism is not the common good. Capitalism has nothing to do with social responsibility. If capitalists express concern with social responsibility, the unified theory (even neoclassical theory) would say that they are seeking profits with guile. According to the unified theory, the same conclusion applies to politicians: they are interested in buying votes and maintain the political power; hence, the common good is subsidiary to that goal. The production and distribution outcome of the capitalist system depends upon the degree of economic and political power concentration.

A high degree of concentration, as that of today, has produced a socially inefficient solution; but a much less concentrated capitalism would produce a solution that is socially superior. This is the problem that the unified theory has discovered using scientific rules.

14.3 Current Public Policies

The capitalist system cannot produce any longer more quantity of goods and higher quality of life at the same time. In addition to the problem of social disorder, the problem of environmental degradation has set limits to better quality of life. Social choices must be made between growth and quality of life at this stage of capitalist development.

A possible procedure would be that social scientists present to society the set of alternative public policies that are based on a valid scientific theory (the unified theory in this case) together with the normative principles that apply to each alternative. However, this is not how public policies work in the real world. Actually, public choices on growth and quality of life have already been made by social actors through the democratic system. The current public policy can be characterized as the pro-growth choice. Certainly, we are living the era of economic growth, with countries racing to win the first places in the track of economic growth. No other criterion to evaluate the economic performance of countries seems to exist now, but economic growth.

How has this social choice been made? Policy alternatives are not socially neutral. Some policies benefit more to some social groups. According to the unified theory, policy makers are not altruists, but are just ordinary people who also act guided by the motivation of self-interest; moreover, democracy is the public choice mechanism, but it includes formal and informal rules (Chapter 7). Hence, there is social conflict in policy choices. The political and economic power, national and international, dominates in the social choice process. Therefore, we could conclude that the pro-growth choice reflects the current power structure, dominated by the current economic and political elites.

In the process of economic growth, the winners have indeed been the economic and political elites. Within-country inequality has not changed much; so these elites have maintained their income shares and privileges in each country. Between-country inequality has increased, as the income level gaps between the First World and the Third World have increased. The losers have been the workers of the Third World.

Many structural changes have taken place in the process of capitalist economic growth, such as sectorial re-composition, regional re-composition, modern forms of production and consumption. In particular, the rate of technological progress has been spectacular. However, power structure has not changed. According to the unified theory, there are no mechanisms embedded in the growth process that can redistribute economic and political assets as growth proceeds; on the contrary, the mechanisms operate in the direction of maintaining the power structure or increasing it. One can then understand why the economic elites have applied pro-growth public policies at the world scale.

How is this public policy choice legitimized in a democratic system? The use of a paradigm seems to be the subtle mechanism. The political and economic elites make people believe that growth is the best choice for society as a whole. In order to present it as “science-based policy,” the appropriate scientific backing is needed, which has to be an economic theory that is consistent with the paradigm.

According to science historian Thomas Kuhn (1970), a theory becomes a paradigm when it is accepted by the community of scientists. In the case of economics, the economic theory paradigm goes beyond the scientific community and expands into the public opinion to become the public policy paradigm. This is feasible to do because the mass media are in

the hands of the elites. Once an economic theory is applied to public choice, it becomes in turn reinforced and legitimized as the paradigm. The equilibrium situation on public policy is thus reached. Therefore, one could say that the economic paradigm utilized in public policies depends upon the distribution of power structure; that is, different power structures will seek to use different economic theories and public policy paradigms. The economic paradigm is thus endogenous to the economic process.

An economic paradigm is not necessarily an economic theory that has survived the scientific process of falsification. In this case, the incentive system to accept a theory is not based on scientific knowledge but on vested interests. In fact, the process of scientific research in economics follows another path. Basic research and applied research (within the paradigm) are mostly independent processes in the real world.

The current public policy paradigm is that economic growth is a necessary and sufficient condition to solve the social problems of individual capitalist countries and those of the world as whole; moreover, capitalism is the only system that can do it. Economic growth with the persistence of high degree of inequality is politically legitimized by applying policies to reduce absolute poverty, which governments use as a mechanism to buy votes. Economic growth thus becomes a paradigm and is believed to be the panacea.

The current pro-growth policy paradigm is based on standard economics. A simple rule to recognize the current economic theory paradigm is to look at the content of the university textbooks in economics. Just like physics, economics is taught at the world scale using textbooks that contain neoclassical and Keynesian theories. Public policy principles are then derived from these theories. The equilibrium situation in public policy choice is thus reached.

Within standard economics, economic growth is the normative objective in the neoclassical theory, which is the dominant theory. Growth is considered to be even more important than short run macroeconomic policies (Keynesian theory). As we can read in a popular textbook on neoclassical growth theory: “growth policy options can contribute much more to improvement in standards of living than has been provided by the entire history of macroeconomic analysis of countercyclical policy and fine-tuning. Economic growth is the part of macroeconomics that really matters.” (Barro & Sala-i-Martin 2004, p.6). Some authors recognize that capitalist societies are not homogeneous, and thus propose “many recipes but one economics” (Rodrik 2007); however, the “one economics” refers to standard economics, which proposes to seek economic growth maximization.

14.4 Alternative Public Policies

Economic growth is not a panacea, as shown by the unified theory. The cost of economic growth is inequality and social disorder, together with the environment degradation, which implies lower quality of life for the present and future generations. More importantly, the problem at hand involves the survival of the human species, as we know it. Therefore, public policy options should be discussed at the world level. For the sake of simplicity in the argument, assume that the entire world is organized in the form of capitalism. So the policy implications of the unified theory can be utilized in a meaningful way.

As pointed out earlier, the current public policy has chosen one extreme in the feasible set of social objectives: to seek economic growth maximization. Because economic growth does not solve the social problems of inequality and quality of life, the alternative public policy consists in reversing the current priority on growth. The alternative involves changing the current public policy: to dethrone the unique social objective of economic growth and to enthrone the objectives of a lower degree of inequality and a better quality of life. The alternative to the current public policy thus entails seeking slower growth rates or even zero-growth rates in per capita income.

The objectives of the alternative public policy can be stated analytically with the help of Figure 12.3, Chapter 12. Consider that the main objective is to delay as much as possible the threshold period at which human life, as we know it, will collapse (period T_p^*). This objective can be attained by lowering the economic growth curve (curve DR should be shifted downward to a flatter curve, say DR') and by shifting outward the constraint of non-renewable natural resources (curve BN should be shifted outward, say to another curve BN'). The critical threshold period given by the depletion limit (T_d^*) would clearly be shifted outwards. The critical threshold period given by the pollution limit (T_p^*) would also be shifted outwards. This is due to the fact that lower output per capita implies lower total output, which in turn implies lower pollution; therefore, the critical period of pollution will appear much later. Which critical period will come first is undetermined.

If the policy objective were zero-growth, would the critical period T^* be infinite? No. At zero-growth alternative, total output will keep growing (at the rate given by the population growth rate); hence the pollution and depletion effects of total output will be in operation and the critical period T^* will be finite. Even if total output were constant (so that output per worker falls), this critical period will be finite. This was clearly shown in Figure 12.2, Chapter 12.

How would a zero-growth world society function? This may sound socially unviable to the reader. However, zero-growth does not mean the renunciation of a better quality of life. For one thing, it means a fall in the negative impact of the environmental degradation upon peoples' life. It does mean changes in the way of life of society, as pointed out by theorists who have studied the functioning of "zero-growth societies" (cf. Olson and Landsberg 1975). As we know, John Stuart Mill in the 1850s had already studied societies in a stationary state: with zero-growth but improvements in technology, ethics, and the art of living. As Daly (1996) has nicely summarized, Mill was advocating for economic *development* without economic *growth*, that is, qualitative improvements without quantitative increase (p.3).

Does zero-growth world society imply zero-growth for the Third World countries? No, it doesn't. Under a situation of zero-growth in the global economy, overall inequality could be reduced by two alternative ways. First, the positive economic growth in the Third World could be offset by a negative rate in the First World. Second, both regions could have zero-growth, but consumption would be redistributed from the First World to the Third World, to the poor masses of the Third World, to be more precise. Both options would imply a redistribution of consumption, one in the form of differences in economic growth rate and the other in the form of direct transfer. Both alternatives would certainly imply a fall in the consumption level of the First World and of the elite groups and middle classes of the Third World countries. But the gain is clear: quality of life would improve for all.

Would redistribution of economic growth rates from the First World to the Third World imply a lower *rate* of environmental degradation? The answer is an empirical question. In poor and rich countries greenhouse gas emissions will increase if income increases. This is the income effect on gas emissions. The redistribution effect assumes total income as given and just sees the effects of income transfers from the rich to the poor. If the consumption basket of a poor country had a lower content of mineral resources per dollar of expenditure than the basket of the rich, the total gas emissions would fall if income were transferred from the rich to the poor; that is, what the rich would reduce in emissions is a higher quantity than what the poor would increase. The global rate of environmental degradation would then depend on another variable: the inequality within current generation. The lower the degree of inequality is, the lower the rate of degradation will be.

Available data indicate that the flow of emissions of CO₂ per unit of output in the First World is not too different from that in the Third World for the period 1980-2006 (UNCTAD 2009, Table 5.1, p. 136). This result refutes the hypothesis known as the Environmental Kuznets Curve, which is an inverted U curve, showing the following relation: increasing emissions per capita as income per capita rises in poor countries and decreasing emissions per capita as income per capita rises in rich countries, which would imply that emissions per unit of output is higher in poor countries than it is in rich countries, which is against facts. Hence, according to these empirical data, the redistribution effect would not seem to be significant. Income redistribution from the First World to the Third World would not increase the degree of degradation of the bio-physical environment.

The consequence of this empirical result is that there is no conflict between the social objectives of reducing current between-country inequality and reducing the rate of environmental degradation. Whether redistribution of income from the wealthy to the poor within Third World countries would have a significant redistribution effect is unknown. On the basis of the available data, we may conclude that a reduction in between-country inequality would not increase between-generation inequality; there is no trade-off between these social objectives.

Dethroning economic growth to low or zero-growth rates and enthroning inequality and quality of life as policy goals would certainly involve qualitative changes on our current lifestyle. Instead of seeking mostly the maximization of private consumption goods, as we do now, society should also seek to improve the quality of life for all. This alternative policy involves reducing superfluous consumption. A significant proportion of current consumption is unnecessary for good quality of life and generates an enormous amount of waste. The fetishism of goods would need to be replaced by the search for the satisfaction of human basic needs, as good quality of life, as the final target.

A car, for example, satisfies the need of transport and other things, but if the goal is to satisfy the need for transportation, other forms of organization, as a good public transport system, can be environmentally more efficient in achieving this end. Health economist Philip Musgrove (2007) has shown that income is not the most important factor in the health well-being in our societies. Even the search for lower inequality can be seen not as an end in itself, but as a means to satisfy the need of social order and thus higher quality of life for all.

Improving the quality of life of society in a context of slow-growth or zero-growth also involves the development of social prevention and protection systems against risk. Under the pro-growth paradigm, the gains in consumption level are subject to the risk of losses due to shocks upon society. These shocks can be endogenous or exogenous to the economic process, to human activity. With the possible exception of earthquakes, most shocks are endogenous, such as shocks associated to climate change (natural disasters, epidemics, and diseases), financial crises, economic recessions, inflation, terrorism, drug wars, organized crime, civil wars, and international wars; shocks from climate change are partly endogenous and partly exogenous, as shown above (Chapter 12). Social protection systems would be needed, which implies more public goods than we consume today (and less private goods to compensate for).

Although the unified theory predicts that endogenous shocks will decrease as societies become less unequal and economic growth rates are reduced, market and state forms of prevention and protection against risks of both types—endogenous and exogenous—would need to be developed. A high quality of life society implies a low risk society.

The environmental distress is nowadays another source of social conflict. Most investment projects that seek to exploit natural resources use local resources—water, land, forest—and are thus resisted by rural communities who fear the loss of their livelihoods. These communities are at the forefront of environmental movements in the world. For them, the environment problem is a matter of survival. They know that when the land is mined and the trees are cut by businesses, their water resources dry up or they lose grazing and farm land. Politically, they have no voice or only limited voice. This social conflict is in accord with the entropic economic process developed in this book: there is no such thing as sustainable economic growth. A change in the power structure of society would lead to public policies that are less pro-growth and more improvement in the quality of life.

The apparatus of the intergenerational consumption frontier shows another very dramatic social conflict: given the stock of mineral resources and the rate of technological progress, economic growth *now* implies a higher degree of inequality between the current generation and future generations. Technological progress that is mineral resource-saving cannot eliminate this social dilemma; it can only make the choice less dramatic. In the production and consumption process, the quantitative and qualitative degradation of the biophysical environment will increase continuously and irrevocably. The only choice society has now is on the velocity of the degradation, as shown in Figure 12.2, Chapter 12. Overall, the public policy dilemma consists in making social choices under the constraint and trade-offs given by the intergenerational consumption frontier.

According to the evolutionary model of the unified theory, the main social choices include the following:

- (1) At what rates to grow? Positive or zero global growth rates? And what rates for the First World and what for the Third World?
- (2) We are somewhere near the critical period T^* , which implies that society should intervene more vigorously on environmental policies. Which *new* global policies would those be?

- (3) How much investment should be allocated to technological progress that is both mineral resource-saving and risk-reducing?
- (4) What types of institutional innovations are needed to have higher quality of life under environmental distress?

These public choices will have a definite effect on the degree of inequality between generations (current and future), as they imply changes in the intergenerational consumption frontier. Regarding public choice problem (1), a positive economic growth now is a social choice in favor of the present generation and against the future generations, as shown above (Chapter 12).

In regards to the social choice problem (2), delaying the period of intervention implies less mineral resources left for the future generations and less consumption levels for the future generations. The failure of Climate Conferences to achieve an international commitment on gas emissions reduction is a social choice in the direction against future generations.

As to problem (3), investment in new production technologies that are mineral resource saving in a significant magnitude will improve the terms of the trade-off. However, the international literature indicates mixed results as to the incentives of the capitalists to generate and use green technologies (Greenhalgh 2005, Acemoglu et al 2009). From the quality of life point of view, it is unclear whether new technologies are risk-reducing. Some examples in the real world indicate that new technologies may be rather risk-increasing, such as the observed cases of oil spillover accidents and the radiation from nuclear power plants accidents. Profit maximization rationality and market competition seems to induce firms to put into use not fully tested technologies; firms calculate the internal losses that they can absorb, but no calculations are made for the damages to the environment in case of failure.

Finally, on problem (4), new consumption technologies can also be developed to save mineral resources. The quantity of mineral resources required per unit of consumption should be reduced. This can be attained with technological innovations that reduce this coefficient in the production process and also in the consumption process. Technological innovations have two components: innovations in production technologies and in consumption technologies. Hence social innovations in our current life style that lead to saving mineral resources and reduce waste-pollution are part of public choice.

According to the unified theory, a capitalist society with more equality and better quality of life can be achieved without growth. Economic growth is not a necessary condition for building such a society for the present and future generations. As we have shown above, economic growth has not helped to attain a better society for all; hence, economic growth is not a sufficient condition either.

Economic growth can be subject to quantitative regulation by setting limits to emissions of greenhouse gases. The effect would be to extend the survival period of the human species (period T^*) and, at the same time, to increase quality of life. Technological innovations that are mineral-resource saving, sun energy-augmenting, and risk-reducing would also have these two effects. These changes have not taken place in the period of rapid economic growth because the economic and political elites lack the incentives to do that. The dynamic equilibrium of the capitalist system would remain unchanged as long as

the exogenous variables remain unchanged. According to the unified theory, the most significant exogenous variable is the inequality in the distribution of economic and political assets in the capitalist system. Alternative policies to the current pro-growth policy thus require a change in the power structure of society.

In light of the unified theory, public policy options are clear, and can be summarized as follows. We either maintain the current pro-growth policies, which lead to increases in inequality, environmental degradation, and risks, with the consequent deterioration in the quality of life, or we use scientific knowledge to change the exogenous variable to develop a new social world in which the quality of life is better for the present generation and also for future generations.

14.5 Breaking with History

The exogenous variables of a valid economic theory constitute instruments of public policies. In the case of the unified theory, the exogenous variables are technological progress and the degree of inequality in the distribution of the economic and political assets or the power structure. But technological change is eventually dependent upon the power structure, as shown in the previous section; so we are left with power structure as the only exogenous variable. As long as the power structure remains unchanged, the current path of growth, inequality, and environment will also remain unchanged; that is, to change the latter implies changing the former.

How can the power structure be changed? How could the legacy of the initial inequality be eliminated or weakened? How to break with history?

The unified theory shows that omega societies can become epsilon societies, but sigma societies cannot. They differ only in the inequality in economic assets and political entitlements, not in overpopulation. Socially hierarchical societies (sigma societies) refer to the inequality in the endowment of political assets; these are societies with different classes of citizenship, which in turn is part of the colonial legacy. Therefore, according to the unified theory, the particular history that matters to understand the capitalist system is the colonial history of capitalist countries. The capitalist system can also be seen as a sigma society, as shown in Chapter 13.

According to the unified theory if Third World countries had initiated the process of capitalist development as socially homogeneous, as the First World did, they would have converged to the First World level. In the long run, overpopulation—the other initial condition—is not the significant because, under that condition, it can be eliminated endogenously in the growth process. This is what the omega theory shows. The fundamental problem in the large majority of Third World countries is colonial legacy: these countries initiated capitalist development as socially hierarchical societies.

In order to clarify the meaning of “break with history,” a summary of this history is in order. The colonial history of Third World countries is not uniform. The colonial systems of England, Portugal, and Spain in America correspond to XVI-XVIII centuries; those of England and France, and in minor degree Belgium, Germany, Holland and Italy, in Africa and Asia corresponds to the last third of XIX and part of the XX centuries. Yet, the colonial legacy is uniform: It is a fact that the Third World countries have had a colonial

past and are poorer and more unequal relative to the First World countries, which have never been colonized.

The 23 countries that have been classified as the First World were never under colonial rule (some were colonial powers), except for the United States, Canada, Australia, and New Zealand. In fact these countries were not colonies, but settler territories. The analytical distinction between *colonial system* and *settler territory* is based on the criterion of population density of territories under domination. The population density in the settler territories was relatively low compared to that of the colonial systems, such as Spanish colonies, as economic historians have shown (Engerman & Sokoloff 1997).

Historians have also shown that, around 1825, at the beginning of democratic capitalism in the American continent, the white population represented 18% of total population in Spanish America, 23% in Portuguese Brazil, but 80% in the United States and Canada. The United States and Canada were not colonial systems based on the exploitation of the aboriginal population, but employed Africans as slaves as the colonial systems did. Around 1825, the proportions of black populations in those territories were 23%, 56%, and 17%, respectively; that is, black population was a minority ethnic group in the United States and Canada compared to the case of Brazil (Engerman & Sokoloff, 1997). Therefore, the initial conditions of the United States, Canada, Australia, and New Zealand were more of an omega rather than a sigma society.

In a sample of nearly 80 countries with colonial and settler histories, it was found a negative correlation between their per capita income in 1995 and the population density (population divided by arable land) in their territory in 1500 (Acemoglu, Johnson, & Robinson 2001, Figure 2, p. 1248). Population density was also important to determine the initial inequality and the initial institutional rules. Colonizers concentrated land and political power in a degree that depended on whether they were acting on territories of high or low population density, colonial or settler systems.

The inequality of citizenship in the Third World is a colonial legacy, the result of the persistence of some colonial institutions. Historians usually point out this legacy by indicating the nature of the colonial institutions. It is worth repeating what Dutch historian H.L. Wesseling (2004) stated: "Colonial societies were generally characterized by apartheid and segregation and often were based on notions of innate racial inequalities" (pp. 242-243). Breaking with history then has a well-defined meaning: breaking with colonial legacies, mainly the legacy of citizenship inequality within Third World countries and between the First World and the Third World. This is just what the unified theory implies as change in the initial inequality.

The independence from colonial powers did not imply breaking with colonial rules. It was not a re-foundational shock. A summary of the general view of historians of Latin America has been put as follows: "Latin America is not postcolonial in any meaningful sense. Its nineteenth-century revolutions were predominantly civil wars between creoles (individuals of European descent born in the Americas) and their 'peninsular' (Iberian) cousins. For the non-white majorities, little of import has changed since colonization" (Graubart 2004, p.217).

The essential role played by the colonial history of capitalist countries upon their development process is, however, challenged by some historians. They think that comparing countries that were never colonized with others that were, may help us in

assessing the effect of colonial legacy on the economic development of countries. For example, historian Wesseling says, “If we take a dispassionate look at the facts, it is difficult to regard India and Indonesia, Nigeria, and Egypt—areas that were long and thoroughly colonized and dominated—as less developed than Ethiopia and Afghanistan. Taiwan and Korea, which were for a long time Japanese colonies, are now industrial countries whose economies belong to the fastest growing in the world” (Wesseling 2004, p. 243).

Weak and strong colonial legacies as analytical categories are important to respond to this argument. Taiwan and Korea were Japanese colonies and for a relatively short period only (1910-1945). In addition, one could put forward the hypothesis that the legacy of Japanese colonies is different from that of European colonies. For one thing, this colonial system was not based “on notions of racial inequality”. China, Korea, and Japan are not racially too different, as is the case between European whites and American darks, African blacks, or Asian yellows. Culturally, Koreans speak Korean today, not Japanese; Taiwanese speak their native Chinese, not Japanese. In contrast, Europeans left the legacy of European languages in their colonies. Under these conditions, Japanese colonialism could not generate z-populations, as European colonialism did.

Statistically *the group* of Third World countries shows *on average* a higher degree of colonialism and, at the same time, also shows *on average* a lower income levels and a higher degree of income inequality compared to *the group* of First World countries. This is precisely what the unified theory predicts. However, as in any theoretical proposition, there are exceptions to this general relation. Within the Third World group there are five countries that never were colonies: Afghanistan, Iran, Ethiopia, Thailand, and Turkey (Wesseling 2004). They would be classified as omega societies, yet they *are not* catching up. These five countries could constitute exceptions to the relation predicted by the unified theory. If we take into account the fact that these five countries show high degrees of inequality, in the range of 0.43 to 0.50 (Deininger and Squire (1996), which is comparable to the rest of the Third World, the paradox tends to disappear. Their initial high degrees of inequalities may have had other causes, not colonialism.

As shown by economic historian Angus Maddison (1995), Japan is the only historical case of catching up. A country that in 1820 was among the group of poor countries, Japan has become a full member of the club of the First World countries. The other possible cases for the near future include only South Korea and Taiwan. In light of the unified theory, these three Asian countries started capitalist development as omega societies. In light of the unified theory, these countries’ economic performance cannot be seen as “miracles,” as they are usually called, for the theory predicts this behavior.

The analytical distinction between formal institutions and their degree of enforcement made in the unified theory would prove helpful now. Formally, most capitalist countries operate under the same institutions: private property, market system, and democracy. It is true that some colonial institutions persisted after independence, such as the case of forced labor and slavery; but today capitalist countries operate predominantly with labor markets, a capitalist institution. However, we have to consider the significance of the enforcement of the formal rule of law and the existence of informal rules in the workings of institutions. It is the enforcement of institutions that is relevant for economics.

The unified theory says that the workings of institutions are endogenous and depend upon the degree of inequality in the asset endowments among individuals, including political power under the category of assets. The fundamental colonial institution that persists today is the inequality in citizenship, in spite of what the formal democracy rules may say about citizenship equality within countries or as universal rights.

According to the unified theory, history matters in the process of capitalist development. This prediction of the theory is consistent with facts. Most post-colonial countries are relatively poor and the relatively poor countries are mostly post-colonial countries. Economic development implies path dependence. However, the existence of path dependence does not mean historical determinism. Economics is not physics. Public policies can be utilized to break with history. This implies the introduction of institutional innovations in the democratic system to eliminate the citizenship inequality—the legacy of the colonial history—within the Third World countries and between the Third World and the First World.

Changing the current power structure to a more equal society would imply a higher degree of democracy. The enforcement degree of the current democratic rules would be higher. New democratic rules for the common good would be created. Limits to concentration in the ownership of land, physical capital, and human capital could be included in the new rules. In general, social choices would be different to those taken now. Public policies would give less weight to growth and more to quality of life. Technological progress could be planned for the common good. Technological progress that is mineral saving would be given priority on the research agenda. Technologies for better quality of health and education services would also be given higher priorities.

A truly democratic government at the world scale is also needed because the environmental problems involve the world society, taken as a whole. The democratic innovation includes the idea that world democracy should be based on a new power structure, not on the continuation of the current power structure, which is very concentrated in the economic and political elites. The world democratic government would then imply one world society, with one class of citizens. In principle, this new democratic government could implement public policies to supply those fundamental public goods for better quality of life for all. This world government could then regulate world society.

According to the unified theory, individual behavior of all social actors—workers, capitalists, and politicians—is guided by the motivation of self-interest, which cannot lead in the aggregate to the common good; therefore, individual behavior needs to be effectively regulated for the common good. We live in a world of free-riders. In such a world, the production of public goods runs the risk of not being produced or produced in too small quantities. The fundamental social problems of our time are precisely how to increase the supply of public goods, such as social order, and quality of life for all. An institutional innovation in the democratic system, which can revolutionize the workings of democracy, is then needed for these tasks. The institutional innovation consists in the equalization of citizenship at the world scale.

The public policy implications of standard economics constitute the engineering of the science of economics today. The public policy implication of standard economics calls for economic growth. Economic growth is seen as sufficient condition to solve the social problems of present and future generations. Therefore, the engineering of standard

economics deals with the problems of how to generate higher growth rates of economic growth. Complementary policies refer to how to reduce poverty and how to calculate environmental costs to achieve “sustainable economic growth.” Hence the engineering of standard economics cannot deal with the problems of how to generate social progress for present and future generations without economic growth.

In contrast, the implications of the unified theory include the possibility of social progress without economic growth. These views are new, which are leading to Modern Economics, calling for new paradigm. This is to say that the current paradigm is based upon Old Economics or Standard Economics. Therefore the engineering of the unified theory is something that needs to be developed. This will not be an easy task. We have been living in the growth paradigm for many years and we are used to think in that way only. Breaking an economic paradigm is like changing from a known geometry to another that is unknown.

14.6 Conclusions

The dynamic model of the unified theory has found that the ultimate factor that explains the existence and persistence of the overall income inequality in the capitalist system is the initial inequality in the individual distribution of economic and political assets. Given their different initial inequality, Third World and First World countries have followed different paths of growth and income distribution. The model predicts that economic growth cannot endogenously change the initial inequality; thus within-country and between-country inequalities are persistent, as empirical data show. In the process of capitalist development, there is path dependence; that is, history matters.

The evolutionary model of the unified theory in turn has shown that the continuous and irrevocable degradation of the physical environment, which would occur in a static economic process, has been accelerated with economic growth. Hence, there is no such thing as sustainable economic growth. The side effects of economic growth include the risk of lower quality of life for current and future generations together with another risk: the end of the human species, as we know it, in the finite future. The model thus generates a trade-off between economic growth and quality of life for present and future generations. In this evolutionary model, the exogenous variables are the inequality in asset distribution and the rates of technological progress.

Economic growth is the objective of public policies in the capitalist world of today. According to the unified theory, this is a reflection of the inequality in the distribution of economic and political assets, which determines the current power structure in society. As long as the current power structure remains unchanged, pro-growth public policies will also continue.

Alternative public policies have also been derived from the unified theory. Another dynamic equilibrium path in the capitalist system, which seeks less growth, more equality, less rapid degradation of the environment, and better quality of life, is feasible. But its implementation would require the elimination or significant reduction of the current world power structure. The asset that is crucial in the current power structure is the inequality in the distribution of political assets—the existence of citizenship classes—within the Third World countries and between the First World and the Third World. This is the legacy of the

colonial history of capitalist countries. The public policy would consist in the introduction of an institutional innovation that seeks the equalization of citizenship: one capitalist world with one citizenship class. This would be a way to break with history.

In this new world, democracy would work better and public choices would reflect the interest of the people and would seek the common good. A new democratic system would produce different outcomes by using new public policies. In this new democratic system, plutocracy would be transformed into real democracy—the power of one class of citizens—, in which politicians will now have the incentives to seek the common good, under short run and long run views. The enforcement degree of the current democratic rules would be higher. New democratic rules for the common good would be created. Limits to concentration in capital ownership could be included in the new rules. In general, social choices would be different to those taken now. Public policies would give less weight to growth and more to quality of life. Technological progress could be planned for the common good. This is what the unified theory predicts.

According to the unified theory, individual behavior of all social actors—workers, capitalists, and politicians—is based on the motivation of self-interest, which cannot lead in the aggregate to the common good; therefore, the behavior of all social actors needs to be effectively regulated for the common good. Because we live in a world of free-riders, the production of public goods runs the risk of not being produced or produced in too small quantities. The fundamental social problems of our time are precisely how to increase the supply of public goods, such as equality, social order, and quality of life for all.

A global democratic government could be created as part of the institutional innovations. This is needed to produce the global public goods that are necessary for a better quality of life for all, including the global control of the environmental degradation. Because economic elites operate at the international level through big corporations, while political elites do mostly at the national level, the local political power is generally subordinated to economic power. Capitalists would cut investment in a country that is not friendly; this is the so-called “Kaleckian threat,” a mechanism that disciplines governments, which today seems more significant than ever. The idea with the institutional innovation is that the new power of democracy should prevail over the power of the capitalist class; moreover, it should prevail over the political class, and even over workers, so as to reach the common good.

An institutional innovation in the democratic system, which can revolutionize the workings of democracy, is then needed for these tasks. The needed institutional innovation consists in the equalization of citizenship at national and international levels.

In sum, power structure is the exogenous variable of the unified theory. The growth and quality of life outcome from the economic process will continue over time as long as this variable remains unchanged. A new outcome will require changing this variable. But how does this exogenous variable get changed? What are the factors than can change it? The unified theory has no answer. The reason has to do with epistemology: an economic theory leaves its exogenous variables unexplained. Contrary to physics, economics cannot explain everything because of the existence of endogenous and exogenous variables in the economic process. A *Theory of Everything* is logically impossible in economics. The implication is that the solution to the economic problem lies outside economics. This is another case of the Gödelian Principle of Incompleteness of any logical system (such as a

scientific theory), which in the case of economics is due to the logic of using endogenous and exogenous variables. Exogenous variables come from outside the economic process.

Appendix A

Comparative Statics: Mathematical Proofs

The core of the static general equilibrium models of epsilon, omega, and sigma theories refer to the capitalist sector only. The core contains two equations and two endogenous variables: the nominal exchange rate (P_e) and the quantity of labor demanded (D_h). They were represented by equations (4.12) and (5.7) in the text, Chapters 4 and 5. These equations are identical. Comparative statics were derived graphically in the text, now the mathematical proof is presented in this Appendix. The equations are reproduced here just for convenience:

$$D_h = H(P_e; P_h, z^*, K_b), H_1 > 0, H_2 < 0, H_3 > 0, H_4 > 0$$

$$S_m = M(P_e, D_h; P_h, P_b^*), \text{ where } M_i > 0, \text{ all } i$$

$$D_h < D_h^* < S_h$$

In order to derive beta prepositions, we can follow the standard mathematical procedure. Partial differentiation of both equations with respect to each exogenous variable are applied and the system of equations so constructed is solved by using Cramer's rules.

Effect of changes in z^* :

$$\frac{dD_h^*}{dz^*} = H_1 \frac{dP_e^*}{dz^*} + H_3$$

$$\frac{dS_m^*}{dz^*} = M_1 \frac{dP_e^*}{dz^*} + M_2 \frac{dD_h^*}{dz^*}$$

$$H_1 \frac{dP_e^*}{dz^*} - \frac{dD_h^*}{dz^*} = -H_3$$

$$M_1 \frac{dP_e^*}{dz^*} + M_2 \frac{dD_h^*}{dz^*} = 0$$

$$\frac{dP_e^*}{dz^*} = \frac{\begin{vmatrix} -H_3 & -1 \\ 0 & M_2 \end{vmatrix}}{\begin{vmatrix} H_1 & -1 \\ M_1 & M_2 \end{vmatrix}} = \frac{-H_3 M_2}{H_1 M_2 + M_1} = \frac{-(+)(+)}{(+)(+) + (+)} = (-)$$

$$\frac{dD_h^*}{dz^*} = \frac{\begin{vmatrix} H_1 & -H_3 \\ M_1 & 0 \end{vmatrix}}{\begin{vmatrix} H_1 & -1 \\ M_1 & M_2 \end{vmatrix}} = \frac{H_3 M_1}{H_1 M_2 + M_1} = \frac{(+)(+)}{(+)(+) + (+)} = (+)$$

Effect of changes in S_m :

$$\frac{dD_h^\circ}{dS_m} = H_1 \frac{dP_\varepsilon^\circ}{dS_m}$$

$$\frac{dS_m^\circ}{dS_m} = M_1 \frac{dP_\varepsilon^\circ}{dS_m} + M_2 \frac{dD_h^\circ}{dS_m}$$

$$H_1 \frac{dP_\varepsilon^\circ}{dS_m} - \frac{dD_h^\circ}{dS_m} = 0$$

$$M_1 \frac{dP_\varepsilon^\circ}{dS_m} + M_2 \frac{dD_h^\circ}{dS_m} = 1$$

$$\frac{dP_\varepsilon^\circ}{dS_m} = \frac{\begin{vmatrix} 0 & -1 \\ 1 & M_2 \end{vmatrix}}{\begin{vmatrix} H_1 & -1 \\ M_1 & M_2 \end{vmatrix}} = \frac{1}{H_1 M_2 + M_1} = \frac{(+)}{(+)(+) + (+)} = (+)$$

$$\frac{dD_h^\circ}{dS_m} = \frac{\begin{vmatrix} H_1 & 0 \\ M_1 & 1 \end{vmatrix}}{\begin{vmatrix} H_1 & -1 \\ M_1 & M_2 \end{vmatrix}} = \frac{H_1}{H_1 M_2 + M_1} = \frac{(+)}{(+)(+) + (+)} = (+)$$

Effect of changes in K_b :

$$\frac{dD_h^\circ}{dK_b} = H_1 \frac{dP_\varepsilon^\circ}{dK_b} + H_4 \frac{dK_b}{dK_b}$$

$$\frac{dS_m^\circ}{dK_b} = M_1 \frac{dP_\varepsilon^\circ}{dK_b} + M_2 \frac{dD_h^\circ}{dK_b}$$

$$H_1 \frac{dP_\varepsilon^\circ}{dK_b} - \frac{dD_h^\circ}{dK_b} = -H_4$$

$$M_1 \frac{dP_\varepsilon^\circ}{dK_b} + M_2 \frac{dD_h^\circ}{dK_b} = 0$$

$$\frac{dP_\varepsilon^\circ}{dK_b} = \frac{\begin{vmatrix} -H_4 & -1 \\ 0 & M_2 \end{vmatrix}}{\begin{vmatrix} H_1 & -1 \\ M_1 & M_2 \end{vmatrix}} = \frac{-H_4 M_2}{H_1 M_2 + M_1} = \frac{-(+)(+)}{(+)(+) + (+)} = (-)$$

$$\frac{dD_h^\circ}{dK_b} = \frac{\begin{vmatrix} H_1 & -H_4 \\ M_1 & 0 \end{vmatrix}}{\begin{vmatrix} H_1 & -1 \\ M_1 & M_2 \end{vmatrix}} = \frac{H_4 M_1}{H_1 M_2 + M_1} = \frac{(+)(+)}{(+)(+) + (+)} = (+)$$

The effect of changes in each exogenous variable upon the rest of the endogenous variables is obtained just by implication from the solutions of the core. Because epsilon, omega, and sigma static models have different endogenous variables, the effect of each exogenous variable upon the endogenous variables will also be different. In the final static models, the beta propositions regarding total output and degree of inequality in each type of society appear as the sign of the partial derivatives in the reduced form equations (7.5) and (7.6) in Chapter 7.

Appendix B

Regression Analysis on the Stability of Inequality

New empirical evidence on the stability of inequality over time was presented in Chapter 11. The statistical model and the data base utilized are presented in this Appendix.

The main idea is to replicate the statistical analysis made by Li, Squire, and Zou (1996) with a new database provided by Milanovic (2010). The new sample includes 24 countries (16 from the First World, 2 from the Third World with weak colonial legacy, and six from the Third World with strong colonial legacy) that have nine or more observations on Gini coefficients calculated from net income of households in the period 1950-2008. The sample is shown in Table B.1.

The regression equation from Li, Squire and Zou's paper is reproduced here:

$$g_{it} = \phi_i D_i + \theta_i t_i + \delta_1 d_1 + \delta_2 d_2 + \delta_3 d_3 + \omega_{it}$$

In this equation g_{it} is the Gini coefficient, $i=1,2,\dots,N$ (number of countries), $D_i = 1$ for country i and 0 otherwise, $t_i = 1,2,\dots,T_i$, and $\omega_{it} \sim \text{iid}(0, \sigma_\omega^2)$. The panel data are unbalanced since in general $T_i \neq T_j$ for $i \neq j$. In light of the ANOVA results, Li, Squire and Zou use the adjusted data but include definitional dummies to test for any remaining effect. Then d_1 is the control dummy for income (=1)/ expenditure (=0); d_2 is the control dummy for households (=1)/ individual (=0); d_3 is the control dummy for gross income (=1)/net income (=0). Their hypotheses are:

- (a) $H_0^a: \phi_1 = \phi_2 = \dots = \phi_N$,
- (b) $H_0^b: \theta_i = 0$, for $i=1, 2, \dots, N$.

Because the new database from Milanovic uses only net incomes, it is not necessary to use dummies (d_1 , d_2 and d_3). This leaves us with the following equation:

$$g_{it} = \phi_i D_i + \theta_i t_i + \omega_{it}$$

The results of the regression are showed in Table B.2.

In Table B.1, numbers in bold indicate time trend is statistically significant at $p=0.05$. Sign & indicates time trend is statistically significant and also quantitatively important, which is defined as a time trend coefficient that is more than 1% of the Gini coefficient of 1980, which implies approximately 20 years to move the coefficient 5 points (Li et al. 1998, p. 33).

Table B.1. Capitalist Countries: Gini Coefficients based on Household Net Income per Capita, 1950-2008 (Countries with nine or more observations)

COUNTRY	N° OBSERVATIONS	MEAN	STD.DESV	MIN	MAX
AUSTRALIA	18	35.48	4.01	31.70	44.00
BAHAMAS	11	45.24	3.58	40.64	52.30
BANGLADESH	10	36.67	5.20	29.00	49.50
BRAZIL	9	58.14	3.39	53.19	63.66
CANADA	23	32.24	1.35	29.40	36.64
CHILE	12	53.01	3.78	45.64	56.98
DENMARK	23	32.51	5.76	22.92	41.27
FINLAND	14	29.84	5.75	22.44	46.00
FRANCE	12	38.26	7.58	29.70	49.00
GERMANY	22	33.19	4.23	28.13	39.60
IRELAND	9	35.84	1.99	31.95	39.00
ITALY	27	36.08	3.29	30.10	42.00
JAPAN	23	34.82	1.35	32.50	37.60
SOUTHKOREA	11	35.10	2.10	32.65	39.10
MEXICO	14	53.14	4.65	40.26	58.00
NETHERLANDS	23	30.27	3.06	26.66	42.00
NEW ZEALAND	14	35.32	3.66	30.04	43.05
NORWAY	19	31.58	4.21	24.50	37.52
SPAIN	9	34.56	1.63	31.99	37.11
SWEDEN	23	28.47	3.79	22.78	39.00
TAIWAN	33	30.19	1.53	27.70	33.60
UNITED KINGDOM	22	33.56	2.39	28.77	37.60
UNITED STATES	46	35.74	1.86	33.50	40.56
VENEZUELA	9	43.16	2.45	39.42	47.65

Source: Sample selected from the database of Milanovic (2010)

Table B.2. Regression Analysis for the Stability of Gini Coefficients, 1950-2008

Country	Country-specific	t-value	Trend Estimate	t-value
AUSTRALIA	34.46754	24.82	0.0516349	0.84
BAHAMAS	48.53662	27.24	-0.2353428	<u>-2.13</u>
BANGLADESH	33.9272	19.03	0.2225856	1.8
BRAZIL	58.25181	31.27	-0.0056694	-0.07
CANADA	32.01795	25.63	0.0118271	0.21
CHILE	47.64499	22.95	0.3423946	<u>2.83</u>
DENMARK	42.41071	23	-0.3988671	<u>-5.69</u>
FINLAND	39.71015	20.21	-0.4201249	<u>-5.48&</u>
FRANCE	49.66241	29.28	-0.4639121	<u>-7.76&</u>
GERMANY	38.082	27.56	-0.167322	<u>-3.97</u>
IRELAND	37.15419	16.5	-0.0654602	-0.65
ITALY	38.80163	37.57	-0.1752088	<u>-3.15</u>
JAPAN	35.13183	28.97	-0.0232317	-0.3
SOUTH KOREA	35.93119	25.3	-0.0445461	-0.75
MEXICO	52.03109	34.88	0.0453969	0.87
NETHERLANDS	33.40303	17.74	-0.1159351	-1.76
NEW ZEALAND	30.22198	18.77	0.4323516	<u>3.62&</u>
NORWAY	36.76986	24.36	-0.2251831	<u>-3.84</u>
SPAIN	34.04266	17.4	0.0232443	0.31
SWEDEN	33.57429	20.41	-0.2112067	<u>-3.34</u>
TAIWAN	29.7904	26.65	0.019715	0.4
UNITED KINGDOM	32.64395	31.99	0.0559011	1.13
UNITED STATES	33.2839	39.12	0.1023728	<u>3.34</u>
VENEZUELA	42.66938	18.31	0.0287867	0.23
NOB	436		R2	0.9942
DF	436		F-Test	15.28
Groups	24			

BIBLIOGRAPHY

Acemoglu, D., Johnson S., & Robinson, J. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review*, 91 (5), 1369-1401.

Acemoglu, D., Aghion, P., Bursztyn, L., & Hemous, D. (2009). The Environment and Directed Technical Change. *National Bureau of Economic Research*. Working Paper 15451.

Aeschbach-Hertig, W. (2007). Rebuttal of “On Global Forces of Nature Driving the Earth’s Climate. Are Humans Involved?” by L.F. Khilyuk and G.V. Chilingar. *Environmental Geology*, 52, 1007-1009.

Akerlof, G., & Kranton, R. (2000). Economics and Identity. *Quarterly Journal of Economics*, 115(3), 715-753.

Alesina, A. & Perotti, R. (1996). Income Distribution, Political Instability, and Investment. *European Economic Review*, 40, 1203-28.

Atkinson, A. (1996). Seeking to Explain the Distribution of Income. In John Hills (Ed.). *New Inequalities. The Changing Distribution of Income and Wealth in the United Kingdom*. Cambridge, UK: Cambridge University Press.

Atkinson, A., Piketty, T., & Saez, E. (2011). Top Incomes in the Long Run of History. *Journal of Economic Literature*, 49(1), 3–71

Banfield, E.C. (1958). *The Moral Bases of a Backward Society*. New York: The Free Press.

Barro, R. (1997). *Macroeconomics* (5th ed.). Cambridge, Mass.: MIT Press.

Barro, R. & Lee, J. (2000). International Data on Education Attainment. Updates and Implications. *The National Bureau of Economic Research*. Working Paper 7911.

Barro, R. & Sala-i-Martin, X. (2004). *Economic Growth* (2nd Edition). Cambridge, MA: The MIT Press.

Baumgärtner, S. (2004). The Inada conditions for material resource inputs reconsidered. *Environmental and Resource Economics*, 29(3), 307-322.

Blanchard, O. (2009). *Macroeconomics* (5th Ed.). Upper Saddle River, NJ: Prentice Hall.

Blinder, A. (1987). Credit Rationing and Effective Supply Failures. *The Economic Journal*, 97, 327-352.

Boulding, K. (1976). The Great Laws of Change. In Tang A., Westfield, F., and Worley, J. (editors), *Evolution, Welfare, and Time in Economics*. Lexington, Mass.: Lexington Books.

- Bourguignon, F. (2000). Crime, Violence and Inequitable Development. In *Annual World Bank Conference on Development Economics 1999*. Washington DC: The World Bank.
- Bowles, S. (1985). The Production Process in a Competitive Economy: Walrasian, Neo-Hobbesian, and Marxian Models, *American Economic Review*, 75(1), 16-36.
- Chilingar, G.V., Sorokhtin, O.G., &Khilyuk, L.F. (2008). Response to W.Aeschbach-Hertig Rebuttal of “On Global Forces Nature Driving the Earth’s Climate. Are Humans Involved?” *Environmental Geology*, 54, 1567-1572.
- Carbaugh, R. (2011). *International Economics*. (13th edition.) Mason, OH: Cengage Learning.
- Clugston, C.O. (2012). *Scarcity. Humanity’s Final Chapter?* Bradenton, FL: Booklocker.com, Inc.
- Credit Suisse Research Institute. (2011). *Global Wealth Report 2011*. <https://infocus.credit-suisse.com/data>
- Daly, H. (1996). *Beyond Growth: The Economics of Sustainable Development*. Beacon Press, Boston.
- Dalziel, N. (2006). *Historical Atlas of the British Empire*. London: Penguin Books.
- Darity, W. & Nembhard, J. (2000). Racial and Ethnic Economic Inequality: The International Record. *American Economic Review*, 90(2), 308-311.
- Davies, J., Sandstrom, S., Shorrocks, A., & Wolf, E. (2010). The Level and Distribution of Global Household Wealth. *Economic Journal*, 121, 223.254.
- Deininger, K. & Squire, L. (1996). A New Data Set Measuring Inequality. *The World Bank Economic Review*, 10(3), 565-591.
- Deininger, K. & Squire, L. (1998). New Ways of Looking at Old Issues: Inequality and Growth. *Journal of Development Economics*, 57, 259-287.
- Diamond, J. (1999). *Guns, Germs, and Steel.The Fates of Human Societies*. W.W. Norton & Co.
- Dominici, F. et al (2002). Air Pollution and Mortality: Estimating Regional and National Dose-Response Relationships, *Journal of the American Statistical Association*, 97(457), March, 100-111.
- Downs, A. (1957). *An Economic Theory of Democracy*. New York: Harper Row.
- Engerman, S., & Sokoloff, K. (1997). Factor Endowments, Institutions, and Differential Paths of Growth among New World Economies: A View of American Historians of the United States. In Stephen Haber (Ed.), *How Latin America Fell Behind: Essays on the Economic Histories of Brazil and Mexico, 1800-1914*. Stanford University Press.

Etheridge, D.M., Steele, L.P., Langenfelds, R.L., & Francey, R.J. (1996), Natural and Anthropogenic Change in Atmospheric CO₂ over the Last 1000 Years from Air in Antarctic Ice and Firn, *Journal of Geophysical Research*, 101(D2), 4115-4128.

Fajnzylber, P., Lederman, D., & Loayza, N. (2002). What Causes Violent Crime. *European Economic Review*, 46, 1323-1358.

FAO (Food and Agricultural Organization) (2005). *Global Forest Resources Assessment 2005*. Rome.

Figueroa, A. (2010). Is Education Income-Equalizing? Evidence from Peru. *CEPAL Review*, 102, April, 113-133.

----- (2012). Income Inequality and Credit Markets. *CEPAL Review*, 105, April, 37-51.

Freixas, X. & Rochet, J. (2008). *Microeconomics of Banking*. 2nd Edition. Cambridge, MA: The MIT Press.

Fukuyama, F. (2011). Left Out. *The American Interest*, 6(3), January-February Issue.

Galindo, L. & Samaniego, J. (2010). The Economics of Climate Change in Latin America and the Caribbean: Stylized Facts. *CEPAL Review*, 100, April, p. 69-96.

Galor, O. (2011). Inequality, Human Capital Formation and the Process of Development. *NBER Working Paper Series*, Working Paper 17058.

Galor, O. (2011). *Unified Growth Theory*. Princeton, NJ: Princeton University Press.

Galor, O. & Moav, O. (2004). From Physical to Human Capital Accumulation: Inequality and the Process of Development. *The Review of Economic Studies*, 71(4), pp. 1001-1026.

Galor, O. & Zeira, J. (1993). Income Distribution and Macroeconomics. *The Review of Economic Studies*, 60(1), pp. 35-52.

Galton, F. (1869). *Hereditary Genius: An Inquire into It's Laws and Consequences*. London: Macmillan and Co.

Gardner, H. (1999). *Intelligence Reframed: Multiple Intelligences for the 21st Century*. New York: Basic Books.

Garraty, J. (1978). *Unemployment in History*. New York: Harper Colophon Books.

Georgescu-Roegen, N. (1971). *The Entropy Law and the Economic Process*. Cambridge, MA: Harvard University Press.

Glaese, E. et al. (2004). Do Institutions Cause Growth. *Journal of Economic Growth*, 9, September, pp. 271-303.

- Gollin, D. (2002). Getting Income Shares Right. *Journal of Political Economy*, 110(2), 458-474.
- Graubart, K. (2004). Hybrid Thinking: Bringing Postcolonial Theory to Colonial Latin American History. In E. Zein-Elabdin and S. Charusheela (eds.), *Postcolonialism Meets Economics*, London: Routledge.
- Grimaud, A. and Rouge, L. (2005). Pollution, Non-renewable Resources, Innovation, and Growth: Welfare and Economic Policy. *Resource and Energy Economics*, 27 (2), 109-129.
- Greenhalgh, C. (2005). Why does market capitalism fail to deliver a sustainable environment and greater equality in incomes? *Cambridge Journal of Economics* 29(6), 1091-1109.
- Hall, G., & Patrinos, A. (2005). *Indigenous Peoples, Poverty, and Human Development in Latin America*. New York: Palgrave Macmillan.
- Hanley, N., Shogren, J., & White, B. (2001). *Introduction to Environmental Economics*. Oxford: Oxford University Press.
- Hawking, S. (1996). *A Brief History of Time, Updated and Expanded Edition*. New York: Bantam Books.
- Hertz-Picciotto, I. et al. (2007). Early Childhood Lower Respiratory Illness and Air Pollution. *Environment Health Perspectives*, 115(10), 1510-1518.
- Hirschman, A. (1973). The Changing Tolerance for Income Inequality in the Course of Economic Development. *Quarterly Journal of Economics*, 87(4), 544-566.
- Hobsbawm, E. (2002). *Interesting Times*. London: Abacus.
- Hofman, A. (2000). Standardized Capital Stock Estimates in Latin America: A 1950-94 Update. *Cambridge Journal of Economics*, 24, 45-86.
- Hudson, R.A. (1996). *Sociolinguistics*. Cambridge, UK: Cambridge University Press.
- Huntington, S. (1996). *The Clashes of Civilizations and the Remaking of the World Order*. New York: Simon & Schuster.
- ILO (International Labor Office). (2000). *Panorama Laboral*. Lima.
- ILO (International Labor Office). (2002). *Women and Men in the Informal Economy: A Statistical Picture*. Geneva, Switzerland: Author.
- IMF (International Monetary Fund). (2011). *World Economic Outlook*. April.
- IPCC (Intergovernmental Panel on Climate Change). (2007). *The Physical Science Basis*. Geneva: IPCC Secretariat.
- Jones, C. (1998). *Introduction to Economic Growth*. Norton.

Karl Marx (1867). *Capital*. London, UK: Allen and Unwin, 1938.

Kalecki, M. (1971). *Selected Essays on The Dynamics of the Capitalist Economy 1933-1970*. Cambridge, UK: Cambridge University Press.

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money*. London: Macmillan Cambridge University Press.

Krugman, P. & Obstfeld, M. (2009). *International Economics: Theory and Policy* (8th Edition). New York, NY: Pearson.

Krugman, P. & Wells, R. (2006). *Macroeconomics*. New York, NY: Worth Publishers.

Kuhn, T. (1970). *The Structure of Scientific Revolutions* (2nd Edition). Chicago: University of Chicago Press.

Lafforgue, G. (2008). Stochastic technical change, non-renewable resource and optimal sustainable growth. *Resource and Energy Economics* 30(4), 540-554.

Lewis, A. (1954). Economic Development with Unlimited Supplies of Labor. *The Manchester School of Economic and Social Studies*, 28, 139-191.

Li H, Squire L, Zou H.F. (1998). Explaining international and intertemporal variations in income inequality. *Economic Journal* 108(1), 26-43.

Lucas, R. (1990). Why Doesn't Capital Flow from Rich to Poor Countries. *American Economic Review*, 80(2), 92-06.

Mc Connell, R. et al. (2010). Childhood Incident Asthma and Traffic-Related Air Pollution at Home and School. *Environment Health Perspectives*, 118(7), 1021-1026.

MacFarling Meure, C. et al. (2006). Law Dome CO₂, CH₄, N₂O Ice Core Records Extended to 2000 years BP. *Geophysical Research Letters*, 33, LI4810.

Maddison, A. (1995). *Monitoring the World Economy, 1820-1992*. Paris: Development Centre, OECD.

----- (2003). *The World Economy: Historical Statistics*. Paris: OECD.

Markusen, J. (2002). *Multinational Firms and the Theory of International Trade*. MIT Press.

Maslow, A. (1970). *Motivation and Personality*. (2nd Edition), New York, NY: Harper & Row Publishers.

Mayr, E. (1997). *This is Biology*. Harvard University Press.

McEvedy, C. (1972). *The Penguin Atlas of Modern History (to 1815)*. London: Penguin Books.

Meadows, D. et al. (1972). *The Limits to Growth: A Report to the Club of Rome's Project on the Predicament of Mankind*. New York, NY: Universe Books.

Milanovic, B. (2005). *Worlds Apart: Measuring International and Global Inequality*. Princeton, NJ: Princeton University Press.

-----, (2010). All the Ginis Dataset. Web Page at World Bank: <http://go.worldbank.org/9VCQW66LA0>. Last Date accessed: 18/10/2010

Mohai, P., Kweon, B.S., Lee, S., & Ard, K. (2011). Air Pollution around Schools is Linked to Poorer Student Health and Academic Performance. *Health Affairs*, 30(5), 852-862.

Mukhopadhyaya, K. & Forssell, O. (2005). An Empirical Investigation of Air Pollution from Fossil Fuel Combustion and its Impact on Health in India during 1973–1974 to 1996–1997. *Ecological Economics*, 55 (2), 235–250.

Muller, E. (1997). Economic Determinants of Democracy. In M. Midlarsky (Ed.), *Inequality, Democracy, and Economic Development*. Cambridge, UK: Cambridge University press.

Muller, R. (2008). *Physics for Future Presidents*. New York: Norton.

Musgrove, P. (2007). The Dethronement of Income as a Cause of Health: an Essay. *Brazilian Journal of Mother and Child Health*, 7(4), 461-466.

Nolan, P. (2006). Marx, Karl (1818-1883). Clark, D. A. (editor), *The Elgar Companion of Development Studies*. Cheltenham, UK: Edward Elgar.

North, D. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge University Press.

OECD (Organization for Economic Cooperation and Development) (2010). *PISA 2009. Results: Executive Summary*. www.pisa.oecd.org.

Okun, A. (1975). *Equality and Efficiency: The Big Trade Off*. Washington, DC: The Brookings Institution.

Olson, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, Mass.: Harvard University Press.

Olson, M., & Landsberg, H. (Ed.) (1975). *The No-growth Society*. London and New York: Frank Cass.

Orchard, L., & Stretton, H. (1997). Public Choice: Critical Survey. *Cambridge Journal of Economics*, 21, 409-430.

Parker, J, Akinbami, L., & Woodruff, T. (2009). Air Pollution and Childhood Respiratory Allergies. *Environment Health Perspectives*, 117(1), 140-147.

Persson, T. & Tabellini, G. (2000). *Political Economics. Explaining Economic Policy*. Cambridge, MA: MIT Press.

Popper, K. (1968). *The Logic of Scientific Discovery*. London, UK: Routledge.

Popper, K. (1985). The Rationality Principle. D. Miller (editor), *Popper Selections*. Princeton, NJ: Princeton University Press.

Psacharopoulos, G., & Patrinos, H. (1994). *Indigenous People and Poverty in Latin America. An Empirical Analysis*. Aldershot: Avebury.

Rand, J., & Tarp, F. (2002). Business Cycles in Developing Countries: Are They Different? *World Development*, 30(12), 2071-2088.

Ratey, J. (2002). *A User's Guide to the Brain*. Vintage Books.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Ricardo, D. (1821). *On the Principles of Political Economy and Taxation*. Cambridge, UK: Cambridge University Press, 1951.

Rodrik, D. (2007). *One Economics, Many Recipes. Globalization, Institutions, and Economic Growth*. Princeton, NJ: Princeton University Press.

Roemer, J. (1982). *A General Theory of Exploitation and Class*. Cambridge, MA: Harvard University Press.

Rousseau, J. (1755). *A Discourse on Inequality*. (Translation from French by Maurice Cranston.) New York, NY: Penguin Books, 1984.

Rogoff, K. (1990). Equilibrium Political Budget Cycles. *American Economic Review*, 80(1), 21-36.

Samuelson, P. (1947). *Foundations of Economic Analysis*. New York, NY: Atheneum, 1965.

Schneider, F., & Enste, D. (2000). Shadow Economics: Size, Causes, and Consequences. *Journal of Economic Literature*, 38, 77-114.

Schultz, T (1975). The Value of the Ability to Deal with Disequilibria. *Journal of the Economic Literature*, 13, 827-846.

Searle, J. (1995). *The Construction of Social Reality*. New York: The Free Press.

Shapiro, C. and J. Stiglitz (1984). Equilibrium Unemployment as a Worker Discipline Device. *American Economic Review*, 74(3), 433-444.

Sheffield, P., Roy, A., Wong, K., & Trasande, L. (2011). Fine Particulate Matter Pollution Linked to Respiratory Illness in Infants and Increased Hospital Costs. *Health Affairs*, 30(5), 871-878.

Silva, N. (2001). Race, Poverty, and Social Exclusion in Brazil. In Estanislao Gacitúa, Carlos Sojo, and Shelton Davis (editors), *Social Exclusion and Poverty Reduction in Latin America and the Caribbean*. Washington, DC: The World Bank.

Smith, A. (1776). *An Inquire into the Nature and Causes of the Wealth of Nations*. New York, NY: Random House, 1937.

Solow R (1974). Intergenerational Equity and Exhaustible Resources. *Review of Economic Studies* 4: 29-45.

Stewart, F. (2001). *Horizontal Inequalities: A Neglected Dimension of Development*. In WIDER, Annual Lecture 5 presented at United Nations University, Helsinki.

----- (ed.) (2008). *Horizontal Inequalities and Conflict Understanding. Group Violence in Multiethnic Societies*. New York, NY: Palgrave Macmillan.

Stiglitz, J., & Weiss, A. (1981). Credit Rationing in Markets with Imperfect Information. *American Economic Review*, 71(3), 393-410.

Strauss, M. (2012). Looking Back on the “Limits to Growth.” *Smithsonian Magazine*. April.

Telles, E. (1993). Urban Labor Market Segmentation and Income in Brazil. *Economic Development and Cultural Change*, 41(2), 231-249.

UNCTAD (United Nations Conference on Trade and Development) (2006). *Trade and Development Report*. New York and Geneva: United Nations Publication.

_____(2009). *Trade and Development Report*. New York and Geneva: United Nations Publication.

UNDP (United Nations Development Program).(2004). *Human Development Report 2004*. New York.

Varian, H. (2002). *Intermediate Microeconomics: A Modern Approach*. 6th Edition. New York, NY: N.W. Norton & Company.

Walras, L. (1883). *Elements of Pure Economics*. New York, NY: A. Kelly Publishers, 1954.

Wesseling, H.L. (2004). *The European Colonial Empires 19815-1919*. London: Pearson.

WHO (World Health Organization). (2006). *Quantifying Environmental Health Impact*. Geneva. <http://www.who.org>

World Bank. (2001). *World Development Report 2000/2001: Attacking Poverty*. Washington, DC.

World Bank Development Indicators (2010) <http://data.worldbank.org/data-catalog/world-development-indicators> .

Wright, E. (1997). *Class Counts: Comparative Studies in Class Analysis*. Cambridge, UK: Cambridge University Press.

Yim, S. and Barret, S. (2012). Public Health Impacts of Combustion Emissions in the UK. *Environmental Science and Technology*, 46 (8), 4291-4296.